Form and Function: Assessing the Impact of Mental Representation on Behaviour using Computational Models

by

Jake Spicer

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Psychology

University of Warwick, Department of Psychology

June 6, 2019

Contents

List of Figures

List of Tables

Acknowledgements

I wish to thank my supervisors, Adam Sanborn and Elliot Ludvig, for their constant support and advice in performing this research, our collaborator Ulrik Beierholm for his assistance in our joint work, our MSc student Jessica Nutt for her help in running experiments, and the staff and students of the Psychology department for allowing all of this work to happen.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. All work presented (including data generated and data analysis) was carried out by the author.

The following parts of this thesis have been published by the author: sections of the introduction and conclusion of the thesis were published as a review paper in Current Opinion in Neurobiology. Part of Chapter 2 was published in the proceedings of the Conference for the Cognitive Science Society 2017. An edited form of Chapter 3 was submitted for publication to Cognitive Psychology, and is currently under review.

Abstract

This thesis presents three studies examining the methods used by human learners to construct mental representations to reflect external data patterns, and the impact the form of these representations have on subsequent behaviour. This involves three varied tasks in which representations are built and updated from experience: stereotype change, numerical estimation and learning consolidation. Each of these studies uses computational models of these processes to offer potential descriptions of the mechanisms used to construct our representations, and assesses the accuracy of these descriptions using both qualitative and quantitative comparisons with human behaviour. Such contrasts reveal the importance of the form of our mental representations on related actions: stereotypical beliefs are coloured by the organisation of group members, numerical expectations are dependent on the assumed format of numerical information, and stimulus choices are influenced by connections forged through experience. This then provides insight into the mechanisms used by human learners in these tasks, and the specific impacts of such mechanisms on related behaviour. We do however also note questions raised by the use of such methods on the accuracy of what may be highly-complex systems in describing human behaviour, and the algorithms that may be used to implement such systems in real life.

Chapter 1

Introduction

A common question in the field of cognitive science: with all the many, varied and complicated items and experiences a person encounters in every day life, how does one go about organising this information into a form which is both usable and useful? Despite the extensive variety of real-world events, people display a remarkable ability to acquire complex representational forms such as item taxonomies, latent patterns and causal structures simply through experience. Such learning suggests the use of advanced mental systems to build these representations, identifying such patterns in our observations to ensure an accurate depiction of true external structures. These systems then play an integral role in directing our behaviour, determining the form of our representations and so our understanding of the world. As such, it is crucial to examine how we build these representations, and how this in turn determines our actions.

Many theorists have therefore sought insight into these processes using computational models of behaviour, drawing on an extensive set of methods by which such structures could be generated in artificial agents. These models can take a number of different forms, ranging from systems which organise observed items by similarity to determine underlying structures, to more abstract systems which represent concepts using boundary rules or networks of weighted connections between lower-order elements. This then provides a diverse range of techniques for use in describing the ways in which learners construct their representations, offering

multiple potential approaches for examining our own systems. Contrasts of these methods with human behaviour can then provide greater insight into the mechanisms which support such acquisition in our own learning, and the results this may have on subsequent behaviour.

This thesis therefore begins with a brief overview of such models, noting differences in form and operation, applications to real-world phenomena, and variations in complexity. Due to the extensive range of such approaches, we here focus on three key branches of these systems: spatial methods, logical methods and network methods, noting the more prominent methods within these branches in ascending complexity. This focuses primarily on applications to human categorisation due to the extensive investigation of mental representations in this area, though applications to other tasks are also noted where appropriate. While the following sections do note some differences between these branches in operation and application, this primarily aims to introduce the methods available when examining human cognition, with greater contrasts between these methods being made in the conclusion of the thesis, found in Chapter 5.

1.1 Spatial Methods

Spatial methods here refers to mechanisms which directly organise actual items and experiences, often placing these items in a multidimensional representational space. One simple form this can take is to store all experienced items in a representational space with a similarity gradient around each for use in future prediction or classification. This is exemplified in recent kernel methods, which use a variety of similarity metrics and often remove redundant stored items to provide a more efficient representation (Jäkel, Schölkopf, & Wichmann, 2007); for example, support vector machines use kernel functions to draw decision boundaries between categories (Cristianini & Schölkopf, 2002), often resulting in highly-accurate classification performance (Decoste & Schölkopf, 2002; Razzaghi, Roderick, Safro, & Marko, 2016; Rasmussen, Rieger, & Webster, 2017). Such methods are mirrored

in behavioural models by exemplar representations, which similarly store all items in a multidimensional feature space, using assessments of similarity to these stored items when making new predictions (Jäkel et al., 2007; Nosofsky, 1986). These models are therefore commonly used as a simple representation of item memory (e.g. Brown, Neath, & Chater, 2007; Nosofsky, Sanders, & McDaniel, 2018), as well as a base for more complex learning models (e.g. ALCOVE, Kruschke, 1992), though concerns have been raised regarding the psychological plausibility of exemplar representations (Vanpaemel & Storms, 2008).

Beyond this direct representation of items, spatial methods can also produce a variety of more abstract representations, with some of the more simplistic being those that aggregate sets of items into collected averages, such as k-means clustering and self-organised maps (Kohonen, 2013; Biehl, Hammer, & Villman, 2016). These methods correspond with the use of prototypes in human learning, often contrasted with exemplar formats, which also use aggregates to represent a set; in the case of prototypes, however, this usually involves only a singular average (Reed, 1972), matching with the most basic form of these methods. Prototypes provide an intuitive method of summarising data sets into an easy-to-use form, but in doing so can miss more complex aspects of item representation, including the relations between stimuli (Vanpaemel & Storms, 2008; Nosofsky, 1992), suggesting greater complexity is in fact required.

A more flexible representation is provided by clustering methods, in which items within a set are assigned to subgroups called clusters according to observed similarities. While this can involve a fixed number of clusters as in the above k-means, much machine learning research has investigated non-parametric forms of this representation, in which the number of clusters is flexible and learned from the data. The Dirichlet Process Mixture Model is one of the most notable non-parametric clustering methods, where the number of clusters is potentially infinite, and inferred from patterns among observations (Antoniak, 1974). More recently, the Indian Buffet Process prior has been used in similar systems to provide alternative representations in which cluster assignments are replaced with feature infer-

ences, again being potentially infinite in number (Griffiths & Ghahramani, 2011). In cognitive science, these processes have been most commonly used within Bayesian models of cognition, providing non-parametric probabilistic systems which infer external structures according to both direct observation and prior beliefs (Anderson, 1991; Austerweil & Griffiths, 2013). These clustering models essentially offer an interpolation between the above exemplar and prototype forms, with each cluster acting as a distinct prototype; any created partition therefore falls between these two extremes depending on the number of clusters formed. This is most evident in rational models of categorisation (Anderson, 1991), though similar techniques have been applied to numerosity (Gershman & Niv, 2013; Sanborn & Beierholm, 2016), language segmentation (Goldwater, Griffiths, & Johnson, 2009) and causal inference (Buchsbaum, Griffiths, Plunkett, Gopnik, & Baldwin, 2015). The present thesis in fact employs such clustering methods in subsequent chapters, using these systems to examine the partitioning of social groups (Chapter 2) and perceptual observations (Chapter 3), and the impact of such processes on related judgements.

Recent advancements in these clustering methods have led to increasingly complex representations, including hierarchical systems where clusters can be nested within each other (Blei, Griffiths, & Jordan, 2010), and the CrossCat model, where multiple partitions of the same items can be formed from different feature patterns (Mansinghka et al., 2016). This has similarly led to the development of hierarchical models of human categorisation (Griffiths, Canini, Sanborn, & Navarro, 2007; Heller, Sanborn, & Chater, 2009), as well as structural form models which select not just the organisation of items but also the form of that organisation, considering clusters, trees and chains among others (Kemp & Tenenbaum, 2008; Lake, Lawrence, & Tenenbaum, 2018). Such models provides a substantial level of flexibility in the ultimate representation, but by expanding the number of considered forms in this way, these systems require strong inductive priors to adequately limit the hypothesis space in order to allow efficient learning from limited data.

1.2 Logical Methods

Logical methods define items or concepts using logical statements concerning the features of the target, identifying common elements within a data set that can be distilled into grammatical terms for use in future predictions. This is most clearly demonstrated in inductive logic programming systems (Muggleton et al., 2012), which have been used to generate logical rules for classifications (Katzouris, Artikis, & Paliouras, 2015) and state transitions (Inoue, Ribeiro, & Sakama, 2014). Similar concepts can be observed in rule-based models of human behaviour, commonly used in categorisation as category membership is often defined by similar boundaries in everyday life (Bruner, Goodnow, & Austin, 1956; Shepard, Hovland, & Jenkins, 1961); indeed, Feldman (2000) suggests that categorisation behaviour reflects the use of Boolean logic, with the difficulty in learning a rule being proportional to its Boolean complexity. Models such as RULEX (Nosofsky, Palmeri, & McKinley, 1994) therefore search through stimulus dimensions to find the simplest rule which maximises discriminability, whilst also creating a store of exceptions. Much like the spatial methods above, these have more recently been developed into more advanced probabilistic grammars (e.g. Goodman, Tenenbaum, Feldman, & Griffiths, 2008), allowing for stronger inductive inferences from limited data. This can again lead to a rational model of structure discovery, in which a rule is inferred from observations using priors on individual components to provide a bias toward simplicity, with lower probabilities for more complex, multidimensional rules. Rule-based systems are not purely limited to categorisation, however, with similar methods being applied to the learning of language (Frank & Tenenbaum, 2011) and functions (Lucas, Griffiths, Williams, & Kalish, 2015).

The key advantage of these logical systems is compositionality: individual elements can be combined to create much more complex rules from fairly simplistic building blocks. In addition, grammars provide a modality-independent representation, able to be translated into alternate formats to direct behaviour in tasks beyond those used for initial learning (Erdogan, Yildirim, & Jacobs, 2015). Logical sys-

tems do, however, naturally draw hard boundaries between categories, making it more difficult to account for the graded nature of human category representations (Rosch, 1973). While probabilistic versions of these systems do help to account for this issue (Goodman et al., 2008; Shepard, 1987), such additional flexibility again requires strong inductive priors in order to learn effectively from limited data.

Recent years have also offered a more advanced form of this representation in 'program' models (Ghahramani, 2015; Lake, Salakhutdinov, & Tenenbaum, 2015), which use Bayesian induction to construct complex production procedures from more basic elements. This is suggested to generate broad and rich representations from small data samples, allowing for accurate generalisations from even a single category member (Lake, Salakhutdinov, & Tenenbaum, 2015) and more intuitive and predictable laws in function learning (Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017). Programs do, however, present an especially complex representational form, and as such are more critically in need of adequate biases to match human learning.

1.3 Network Methods

Network methods provide an alternate form of representation using networks of interconnected nodes, with the strength of the connections being adjusted with experience to reproduce external patterns. This representation intuitively provides a closer correspondence between method principles and actual implementation in the brain: connectionist networks offer a simplified emulation of true neural structures, inherently affording such methods a degree of external validity (McClelland et al., 2010). These systems therefore contrast with both spatial and logical models of human cognition in their level of explanation; while the above methods focus on Marr's computational level, network methods are closer to Marr's implementation level (Marr, 1982).

Rather than the strict delineations between methods seen in the above branches, complexity within these networks increases somewhat gradually according to size,

both in terms of breadth and depth. This extends from basic mechanisms like perceptrons, which essentially provide a connectionist implementation of prototypes (Jäkel et al., 2007; Rosenblatt, 1958), to more complex parallel distributed processing systems, expanding the number of nodes and connections to create a more extensive network with a greater representational capacity (Rumelhart, McClelland, & the PDP Research Group, 1986). There are, however, additional complexities in these methods beyond network size, with recurrent and convolutional networks being some of the more notable forms. In addition, recent neural networks have been further expanded to include external memory stores, using these elements to further improve their performance (Graves et al., 2016).

Highly-complex network methods have in fact become increasingly common in recent years in machine learning due to a surge in the use of deep learning systems in various complex tasks; these methods use multiple, hierarchical layers of connections for increasing levels of abstraction (LeCun, Bengio, & Hinton, 2015; Schmidhuber, 2015). Such systems have the advantage of flexibility, providing a single, global system that can be applied fairly readily to multiple fields. Deep learning systems have therefore been successful in finding categorical structures in image recognition (Farabet, Couprie, Najman, & LeCun, 2013; Krizhevsky, Sutskever, & Hinton, 2017) and speech processing (Hinton et al., 2012; Chen & Mak, 2015), as well as matching or exceeding human performance on complex tasks such as playing video and board games (Mnih et al., 2015; Silver et al., 2016).

Within cognitive modelling, simple network models have been commonly used in associative learning theories (e.g. Rescorla & Wagner, 1972), providing an extensive literature using networks often limited to only a few nodes representing basic stimulus features. Indeed, the present thesis uses such simple network models in Chapter 4 to examine learning consolidation, investigating how associations may be forged between features outside of direct learning to better direct subsequent choices. The more complex networks used in deep learning, meanwhile, are still beginning to be applied to behaviour (Lake, Zaremba, Fergus, & Gureckis, 2015; J. Peterson, Abbott, & Griffiths, 2016; Testolin & Zorzi, 2016), creating cognitive

models that can take advantage of the power of such methods. There are, however, concerns whether such applications are truly valid: while deep learning systems demonstrate a similar level of performance to human learning, and use similar representations to those of actual neural systems (Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014), both speed of learning and ease of generalisation are much higher in people than machines (Lake, Salakhutdinov, & Tenenbaum, 2015; Lake, Ullman, Tenenbaum, & Gershman, 2017), potentially indicating some difference in operation. This is further complicated by the opaqueness of such methods, with any generated representation being distributed across a potentially enormous series of connection weights; this can make interpretation of the learned representation difficult, relying more on behavioural predictions than any obvious structure.

1.4 Summary

As the above should illustrate, past research provides an extensive set of methods which can be used to investigate the ways in which we build and update our own mental representations. This thesis therefore presents applications of such methods to three domains of behaviour, using comparisons with computational models to investigate the processes involved in forming our representations, as well as the resulting differences such forms may have on subsequent behaviour in three varied tasks. While these studies each focus on a different subject, all share a common theme regarding the construction of mental representations from experience: the updating of stereotypical beliefs with new evidence; the effect of prior observations on numerical estimates; and the rehearsal of past trials to consolidate stimulus associations. This also involves a common approach using both theoretical and quantitative contrasts of the predictions from cognitive models of these tasks with actual behaviour to assess the accuracy of these systems, and so provide an indication of the mental mechanisms which underlie our actions. These applications then each demonstrate the specific impact of the form of our representations on related behaviour in these subjects, indicating the direct influence of the operations of our

learning systems on our actions. Such contrasts also raise questions on aspects such as the rationality of human behaviour, the algorithms required to implement these potentially highly-complex systems, and the biases these mechanisms may introduce in our decision making, all aspects which will be revisited in greater detail in the conclusion to this thesis. The three following chapters therefore each present one of these applications, while Chapter 5 closes with common themes raised across these studies.

Chapter 2

A Computational Approach to Stereotype Change

A clear goal when constructing a representation of the environment is to accurately capture any patterns in our observations: similarities or recurring elements within our experiences can provide indications of underlying structures such as item taxonomies or common causes which could be highly valuable in directing related behaviour. As a result, the forms of our representations should be sensitive to the patterns of data we observe, leading to different expectations according to differences in our experiences. In this chapter, we examine this process in the domain of stereotype change, where beliefs about a target category are updated following new observations of group members. This then presents a case in which the representation of the category is altered to reflect new evidence, with the specific form of the new representation determining the influence of such data on stereotypical expectations. This draws on previous work where the organisation of counter-stereotypical information is suggested to affect its use in subsequent judgements: certain patterns may lead to mitigation of such data, preventing any change to existing stereotypical beliefs. We therefore here investigate the process by which such organisation is decided, and how this influences related predictions using computational models of categorisation to provide both qualitative and quantitative comparisons with human data.

2.1 Stereotype Change

While stereotypes may commonly be thought of as distorted caricatures of certain social groups, within the domain of cognitive science, a stereotype can be viewed as a set of traits or behaviours which are expected to be displayed (or specifically not displayed) by an individual based on their membership of a given social category (Hilton & von Hippel, 1996): traits *congruent* with the group stereotype hold high levels of expectation, and traits *incongruent* with the group stereotype hold low levels of expectation. This then raises the question of whether these expectations reflect the holder's summed experience with the category: if such beliefs are based on the prevalence of these traits in observations of actual group members, then stereotypes could be termed as a rational incorporation of available data, even if these expectations are not accurate in all real-world cases. We here use the term 'rational' to refer to a system which uses all available relevant information to produce an estimate, in this case being the predicted likelihood of a trait being observed in future group members; this is in contrast to an 'irrational' system in which some data may be ignored, leading to a bias in the resulting estimate. This aspect can be observed in the changes in stereotypical beliefs when exposed to new information: if rational, stereotypes should adjust to reflect this data, even where it opposes existing expectations.

In reality, however, stereotypes have often been found to be resistant to change, with beliefs and expectations regarding a group often persisting even when faced with directly contradictory information (Hilton & von Hippel, 1996). This presents a problem when trying to combat stereotypes underlying prejudice or discrimination through out-group exposure as has often been suggested by theories such as the Contact Hypothesis (Allport, 1954), as there is no assurance that simply demonstrating the inaccuracy of these beliefs will be effective in encouraging revision. It is therefore necessary to examine the processes by which stereotypes are updated with experience, and, in cases of stereotype persistence, determine how counter-stereotypical information may be disregarded in order to develop better

methods to encourage change.

Past research into this field has offered three possible processes of stereotype revision (summarised by Weber & Crocker, 1983): *book-keeping*, in which the stereotype is slowly adjusted with each relevant observation; *conversion*, in which the stereotype can undergo sudden and drastic changes in response to particularly notable contradictory exemplars; and *subtyping*, in which counter-stereotypical evidence is isolated from the rest of the category in a distinct subgroup, ignored when making category judgements. This presents three potential explanations for stereotype persistence: stereotype-incongruent exemplars may be noted via book-keeping but remain out-weighed by prior stereotypical beliefs; these exemplars may not have been sufficiently significant to evoke change via conversion; or these exemplars may have been excluded entirely via subtyping.

This distinction was examined by Weber and Crocker (1983) by manipulating the presentation format of counter-stereotypical evidence in summaries of lawyers: equal amounts of stereotype-incongruent evidence were either concentrated into only a few exemplars, or dispersed across many exemplars (illustrated in Table 2.1). This generates three competing expectations between the three theories presented above: conversion suggests that these concentrated exemplars showing unexpected traits on all dimensions would act as extreme disconfirmers, encouraging greater revision to the stereotype in the concentrated condition. Conversely,

	Condition					
Exemplar	Concentrated			Dispersed		
1	I	I	I	I	N	N
2	I	I	I	C	N	I
3	C	N	N	N	I	N
4	N	N	N	I	C	N
5	N	N	C	N	N	I
6	N	C	N	N	I	C

Table 2.1: Illustration of the concentration design of Weber and Crocker (1983), showing a subset of exemplars from the concentrated and dispersed conditions, where 'C' represents a stereotype-congruent trait, 'I' represents a stereotype-incongruent trait, and 'N' represents a stereotype-neutral trait.

subtyping would suggest that concentrating incongruent evidence should make it easier to isolate, thereby preserving existing stereotypical beliefs, leading to greater revision in the dispersed condition. Book-keeping, meanwhile, focuses only on the amount of data rather than the presentation format, and so suggests no difference between these conditions. Measures of the strength of stereotypical beliefs following exposure to these exemplars were found to be stronger in the concentrated condition, supporting the subtyping model, an effect that has since been replicated in a number of studies (Johnston & Hewstone, 1992; Bott & Murphy, 2007).

These findings then depict stereotype persistence as the result of an irrational process of purposefully disregarding stereotype-incongruent information by using categorisation mechanisms to exclude this data from judgements. The precise systems underlying these effects remain somewhat unclear however: while multiple studies have attempted to investigate the proposed partitioning of social groups suggested by subtyping using various methods, these often draw on indirect measures such as trait likelihood or exemplar typicality, which may not provide accurate indications of categorisation behaviour (Richards & Hewstone, 2001; Queller & Mason, 2008). Conversely, attempts to directly assess categorisations using methods such as sorting tasks are likely to suffer from demand characteristics, meaning any produced groupings may not be representative of unprompted partitioning of the category (Queller & Mason, 2008). This leaves the actual treatment of counter-stereotypical data in such cases uncertain, offering no assurance that such information is in fact being isolated and ignored as subtyping would suggest.

A more direct assessment of categorisation behaviour can however be obtained using comparisons with computational models of categorisation, contrasting measures of the expectation of stereotypical traits from participants with equivalent predictions of stereotypicality generated by potential categorisation systems to determine which offers the most accurate description of the true process. This method has the advantage of both directly examining the impact of proposed categorisation processes on stereotypical beliefs in a transparent manner whilst avoiding demand characteristics that may occur in other potential measures. The use

of such comparisons can therefore indicate whether learners do indeed divide the category into subgroups, and whether such partitioning leads to the exclusion of counter-stereotypical data; if behaviour best corresponds with such an exclusion mechanism, then this would provide evidence of a specific strategy of stereotype preservation. Conversely, if such fits indicate judgements do actually make use of all available information, then subtyping effects could instead be regarded as a natural reaction of general categorisation processes to data patterns which happen to diminish the impact of incongruent information. This would then present subtyping as a more rational process than it might initially seem, being the result of standard non-parametric categorisation systems creating partitions of all available group data; certain patterns may lead to structures which isolate and so inadvertently mitigate incongruent data, while others generating greater integration of congruent and incongruent information could lead to greater stereotype revision more akin to book-keeping. If so, then the stereotype maintenance associated with subtyping could be fought using similarly rational mechanisms to encourage stereotype change. The application of these techniques to subtyping could then provide a valuable window into the operations of this process, revealing the logical systems underlying this apparently illogical behaviour.

The present study therefore presents a computational approach to stereotype use, investigating both the operations underlying such beliefs and their rationality; in the following sections, we develop several candidate models to approximate the existing depictions of stereotype revision, contrast the predictions of these models with participant data to assess their accuracy, and use these findings to offer some insight into the process of stereotype change.

2.2 Model Details

We consider four potential methods by which stereotypes could be updated following new category member observations, drawing on both the theorised processes described above as well as existing computational models of categorisation. These

are here presented in increasing complexity, defining the precise process of each system, and examining their predictions regarding existing depictions of subtyping. For simplicity, all models considered here are stationary, meaning all stored experiences remain available once an assignment has been made.

2.2.1 The Book-Keeping Model

The first of the candidate models aimed to emulate book-keeping, tracking the rate of stereotype-congruent traits within the category, and updating this prediction with each relevant observation; this was therefore named the 'Book-Keeping Model' (BKM). The BKM stores all exemplars in memory and uses the rate of stereotype-congruent features across relevant stereotypical dimensions in this store to generate predictions of the probability of future category members demonstrating similar traits. This measure is based on a Dirichlet-multinomial distribution, in which the rate of congruent traits across relevant dimensions is combined with an additional parameter to represent prior expectations:

$$p(con) = \frac{1}{d} \sum_{i} \frac{n_{c,i} + \alpha_c}{n_{\cdot,i} + \alpha_0} \tag{2.1}$$

where each i is a stereotype-relevant dimension, $n_{c,i}$ is the number of exemplars presenting congruent values on that dimension, d is the number of these dimensions, and $n_{\cdot,i}$ is the total number of exemplars in the category. The parameter α_c meanwhile reflects the prior expectation of the occurrence of congruent values independent of any observations, equal across dimensions, with α_0 being the sum of α values for all possible trait values on a given dimension.

Predictions from the BKM are therefore based solely on the rate of congruent traits in the category, with no effect of the presentation format of this data; as a result, this model is unable to predict the results of Weber and Crocker (1983), instead suggesting no difference according to the concentration of incongruent information. The BKM therefore acts as a baseline, offering a point of comparison for the other candidate models.

2.2.2 The Strong Subtyping Model

The second candidate model reflects an extreme form of subtyping in which highly incongruent exemplars are directly excluded from the category, preventing these exemplars from influencing subsequent decisions. This corresponds with the 'refencing' concept offered by Allport (1954), in which category boundaries are essentially redrawn to remove counter-stereotypical information. This 'Strong Subtyping Model' (SSM) therefore operates in a similar manner to the BKM described above, but with the addition of a gating mechanism to the memory store: following a new observation, the average rate of incongruent traits of the target exemplar is compared against a pre-set incongruency criterion parameter, labelled θ. Exemplars with incongruency rates below this parameter are added to the store and so influence decisions, whilst those exceeding this criterion are excluded.

The SSM therefore explicitly attempts to remove incongruent exemplars from the category to maintain stereotypical beliefs; as such, the model universally predicts subtyping effects, particularly where exemplars display high levels of incongruency, as in the concentration design of Weber and Crocker (1983).

2.2.3 The Restricted Rational Model of Categorisation

As a counterpoint to this extreme subtyping model, the third candidate model made use of existing non-parametric methods to divide the category into multiple subgroups based on observed similarities between exemplars for a more flexible partitioning of category members. The model is then able to select one of these subgroups for use in generating predictions, presenting an alternate method by which data may be excluded from stereotype-related judgements. This drew on the Rational Model of Categorisation (RMC) defined by Anderson (1991); this model provides a fundamental depiction of non-parametric clustering approaches, making it the most appropriate base for the present model. However, whereas the RMC bases predictions on all formed clusters, the present model included a restriction in considered clusters to allow for exclusion of certain category data; this was therefore

labelled the Restricted Rational Model of Categorisation (RRMC).

Following the structure of the RMC, the RRMC assigns exemplars sequentially to a cluster based on similarities in observed features using a Bayesian model to approximate the ideal partition (where partition refers to the collection of clusters within the category at a given time point):

$$p(k|f) = \frac{p(k)p(f|k)}{\sum_k p(k)p(f|k)} \tag{2.2}$$

where k is the cluster and f is the feature set of the exemplar under consideration. This posterior probability is calculated for all existing clusters as well as a new potential cluster to determine assignment; this could either be deterministic according to the maximum posterior or stochastic, though in the present study we focus on stochastic assignment only. Following Anderson (1991), the prior probability was defined as:

$$p(k) = \begin{cases} \dfrac{cn_k}{(1-c)+cn} & \text{if } k \text{ is old} \\ \\ \dfrac{(1-c)}{(1-c)+cn} & \text{if } k \text{ is new} \end{cases} \tag{2.3}$$

where n_k is the number of exemplars in cluster k, n is the total number of members assigned to the partition, and c is a coupling parameter describing the probability of two exemplars being grouped together independent of any observations. Fixing c at 1 then restricts the RRMC to a single cluster, thereby making this model equivalent to the above BKM. As such, the BKM is nested within the RRMC due to the flexibility of this model, allowing the RRMC to produce similar effects to the BKM where exemplars are similarly grouped together.

The likelihood also followed the format of Anderson (1991):

$$p(f|k) = \prod_i p(j_i|k) \tag{2.4}$$

where the exemplar's features are divided into dimensions i holding values j_i. This matches with the definition for the measure of congruency given by Equation 2.1,

here applied to the observed trait j rather than simply congruent traits:

$$p(j_i|k) = \frac{n_{j,i,k} + \alpha_j}{n_{\cdot,i,k} + \alpha_0} \tag{2.5}$$

where $n_{j,i,k}$ is the number of exemplars in cluster k showing trait value j_i on dimension i, $n_{\cdot,i,k}$ is the number of members of cluster k showing any value on dimension i, and α_j reflects the prior expectation of the occurrence of value j on any dimension, while α_0 is the sum of these α values for that dimension.

Once a partition of clusters has been generated, the model is then able to select one of the created subgroups to provide an estimate of stereotypicality within the category, while other clusters are disregarded. This then allows for the exclusion of data as in the SSM above, though in this case the model does not simply remove highly-incongruent exemplars, but instead focuses on the cluster judged to be most representative within the partition, independent of its contents. This was achieved by restricting the clusters considered when making estimates to only the cluster with the highest posterior probability; such a principle was based on the findings of Murphy and Ross (1994) which suggested that participants often only considered the most likely cluster in their estimates rather than all generated clusters. This used the same measure given in Equation 2.1, here based only on the cluster with the highest posterior probability as defined by Equation 2.2.

Due to the flexibility of this method, the RRMC is able to predict different effects depending on data pattern: where exemplars are grouped together, the RRMC suggests book-keeping, as in the BKM above, while segregating incongruent exemplars can lead to subtyping if the incongruent cluster is not selected for estimates. As such, the RRMC should conform to the results of Weber and Crocker (1983), being more likely to disregard incongruent information where this data is more easily isolated in a distinct subgroup.

2.2.4 The Rational Model of Categorisation

Finally, as an extension of the previous model, we also considered a fully rational model which uses the same non-parametric clustering methodology, but includes all generated clusters in its predictions, matching with the RMC defined by Anderson (1991). The RMC uses the same clustering process defined above for the RRMC to generate a partition, but bases predictions of stereotypicality on the average rate of congruent traits in each cluster weighted by the probability of that cluster:

$$p(con) = \sum_k p(k)p(con|k) \tag{2.6}$$

where $p(con|k)$ matches with Equation 2.1, here calculated separately for each cluster. This measure can also take a more specific form where other exemplar features are already available by using the posterior probability of assignment to a cluster given those features:

$$p(con|f) = \sum_k p(k|f)p(con|k) \tag{2.7}$$

where $p(k|f)$ is given by Equation 2.2.

As with the RRMC above, the flexible representation produced by the RMC leads to different predictions according to different data patterns, with the impact of any observations on estimates being determined by its organisation in the partition. Unlike the RRMC, however, this does not involve the exclusion of any data, meaning incongruent information will impact on predictions even if assigned to a distinct cluster. Even so, the RMC is able to produce a subtyping effect in such scenarios due to differences in the influence of prior expectations between larger and smaller clusters: smaller clusters provide less evidence to outweigh prior expectations, here represented by the α parameter. As such, there is less confidence that future members of the incongruent cluster will demonstrate similar trait values, while the larger congruent cluster carries more certainty.

To illustrate, consider a case in which 30 exemplars, 20 congruent and 10 incongruent, are either integrated or segregated, as depicted in Figure 2.1. For

the purposes of this illustration, $\alpha = 1$ for both congruent and incongruent traits, and $c = 1$, meaning no new cluster is considered. The predicted probability of congruency in future category members is then given by Equation 2.6, being the product of the prior probability $p(k)$, based on the proportion of exemplars in each cluster, and the likelihood of congruency in each cluster $p(con|k)$, based on the rate of congruent members in that cluster, modified by the α values:

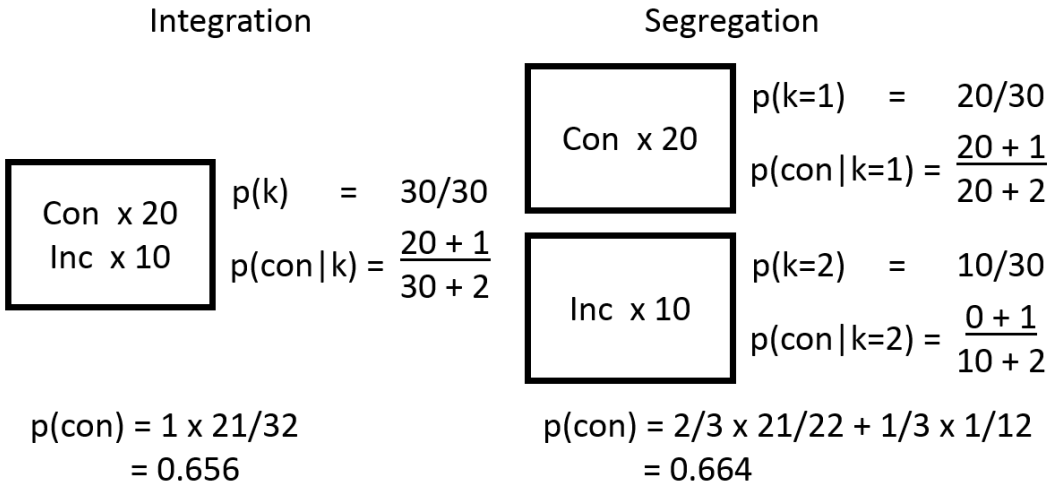

Figure 2.1: A demonstration of subtyping effects within the RMC

This then demonstrates that the α values are more impactful in the smaller cluster, offsetting the actual ratio of traits to a greater degree. As a result, stereotype-congruency is estimated to be more probable in the segregated case, as the smaller, incongruent cluster carries less confidence than the larger, congruent cluster. The RMC is therefore able to produce a subtyping effect without actually ignoring or even down-weighting incongruent evidence: all exemplars contribute an equal amount of information to a prediction, but data patterns in some clusters are more uncertain than others.

2.2.5 Comparing the Models

The four models presented above offer four different mechanisms of stereotype revision: both the RMC and RRMC use a partition that flexibly adapts to observed data patterns, though the RRMC subsequently simplifies this partition by focussing

on only one cluster, heightening any effects this representation may have generated, while the RMC remains more moderate. Conversely, the SSM is a definitive method of stereotype maintenance, with any effect of data pattern being due to the use of a gating mechanism attempting to remove incongruent data from the category. The BKM, meanwhile, focuses on trait ratios rather than data pattern, thereby dismissing any effects that may be predicted by the other candidate models. This then provides three models able to predict some form of a subtyping effect, as observed in Weber and Crocker (1983), and so three potentially valid depictions of stereotype change.

There is, however, a key distinction between these models which can be used to determine their validity: in the RMC, the subtyping effect is dependent on the smaller size of the subtype cluster, meaning that increasing the size of the subtype by adding more incongruent members should reduce and ultimately eliminate this effect. In contrast, the SSM is insensitive to the size of the subtype as these excluded exemplars are no longer considered in the partition; as such, the subtyping effect can only increase with further exposure, as congruent data is accepted and incongruent data is ignored. Similarly, the RRMC will continue to ignore the subtype regardless of its size, though in this case there is a threshold to this process: if the subtype becomes sufficiently large, it may be selected as the most likely cluster, at which point estimates will change drastically to reflect the subtype's much lower rate of congruency. This could essentially reverse the subtyping effect at higher volumes of incongruent evidence, focussing on counter-stereotypical rather than stereotypical clusters, and so bearing a closer resemblance to the conversion-effect described above. The BKM, finally, is unable to exclude incongruent data at all, and therefore predicts no subtyping effect at any volume of incongruent information.

The accuracy of these models can therefore be contrasted according to the change in the subtyping effect with further exposure to stereotype-incongruent evidence: the RMC predicts a reduction in subtyping at higher volumes of counter-stereotypical data; the SSM predicts an increase in subtyping; the RRMC predicts a stable subtyping effect until a sudden reversal; and the BKM predicts no subtyping

effect at any point.

The following sections therefore present three empirical contrasts of these model hypotheses by examining the time course of subtyping in three varying scenarios. These experiments each shared a common base structure, extending the concentration design of Weber and Crocker (1983) across a higher total volume of evidence and taking measures of stereotypical beliefs throughout exposure to look for changes in effect across the task. This also provided direct behavioural data for use in assessing the fit of the candidate models for a more complete test of these predictions.

2.3 Experiment 1

2.3.1 Method

Participants

One-hundred-and-sixteen participants were selected from a University of Warwick undergraduate psychology class as part of a course requirement. The sample included 102 females and 14 males, while age ranged between 18 and 27 years, with a mean of 18.7.

Design and Materials

The experiment followed the concentration design of Weber and Crocker (1983) with an additional within-subjects manipulation of data volume: measures of stereotypical beliefs were taken at fixed intervals during the observation of a set of exemplar descriptions where stereotype-incongruent information was either concentrated in a subset of exemplars or dispersed across all exemplars. Two exemplar sets were therefore created for use in the experiment, each containing 90 total exemplars displaying four trait dimensions: the first dimension described the occupational label, and so was identical for all exemplars, while the remaining three dimensions described personality traits with three possible values (stereotype-congruent,

	Condition					
Exemplar	Concentrated			Dispersed		
1	I	I	I	I	N	I
2	I	I	I	C	I	I
3	I	I	I	I	I	N
4	I	I	I	I	C	I
5	N	N	C	N	I	I
6	N	C	N	I	I	C

Table 2.2: Exemplar structure in the concentrated and dispersed conditions of Experiment 1, where 'C' represents a stereotype-congruent trait, 'I' represents a stereotype-incongruent trait, and 'N' represents a stereotype-neutral trait.

stereotype-incongruent or neutral). In both sets, two-thirds of the 270 total traits were incongruent, one-sixth were congruent and one-sixth were neutral; incongruent traits made up the majority in order to allow for a potential incongruent cluster to be larger than any other in the category. In the concentrated exemplar set, these incongruent traits were concentrated such that 60 exemplars each displayed incongruent traits on all three personality dimensions, with the congruent and neutral traits being distributed equally between the remaining exemplars. In the dispersed exemplar set, all traits were distributed as equally as possible. Exemplar structure in this task is illustrated in Table 2.2.

As in Weber and Crocker (1983), exemplars were said to come from the category of lawyers; exemplars were therefore transformed into member summaries for use in the experiment by assigning each value on the three personality dimensions a unique trait label. Sixteen total labels were used: 5 congruent (Intelligent, Industrious, Neat, Out-going and Well-dressed), 5 incongruent (Incompetent, Lazy, Messy, Shy and Slovenly) and 6 neutral (Warm, Religious, Jovial, Obnoxious, Reserved and Meditative). These labels were taken from Weber and Crocker (1983), being based on pilot tests determining stereotypical and counter-stereotypical traits for the target category of lawyers. In contrast to their use in Weber and Crocker (1983), however, these traits were here presented as discrete labels rather than as descriptive sentences; this was intended to allow for a closer correspondence between participant and model evaluations of exemplar descriptions, keeping trait values discrete

and definitive.

Three labels of each trait type were randomly selected at the start of each run of the experiment for use in exemplar summaries. Summaries were also assigned randomly selected names to assist in individuation.

Procedure

Upon arriving at the lab, participants were first randomly assigned to one of the two concentration conditions, determining which set of exemplars would be viewed; this was balanced to provide equal numbers, meaning 58 participants were allocated to each condition. Participants were told the experiment tested how perceptions of a group changed with experience, involving both viewing summaries of group members and answering questions about the traits of the group in general.

The experiment began by asking participants to estimate the likelihood of certain traits appearing in the category of lawyers. To provide a more intuitive measure of probability, participants gave estimates of the number of members in a sample of 100 lawyers displaying each trait; for example, 'Out of 100 lawyers, how many do you think would be: intelligent?'. Estimates were requested for all 16 possible personality traits, though only 9 were used in the subsequent member summaries. This first question block therefore provided a measure of baseline beliefs before any experimental exemplars were viewed. Figure 2.2a shows a sample slide from this measurement block.

After providing estimates for all traits, participants began a presentation block in which member summaries were shown on screen for the participants to examine. In order to maintain attention on this information, participants were asked to rate the pleasantness of each group member on a scale of 1-10, though this measure was not used during analysis. Figure 2.2b shows a sample slide from the presentation period, including a highly-incongruent exemplar from the concentrated condition.

At set intervals of presentation, the test block was repeated, and participants were again asked to estimate the likelihood of each of the 16 traits appearing in

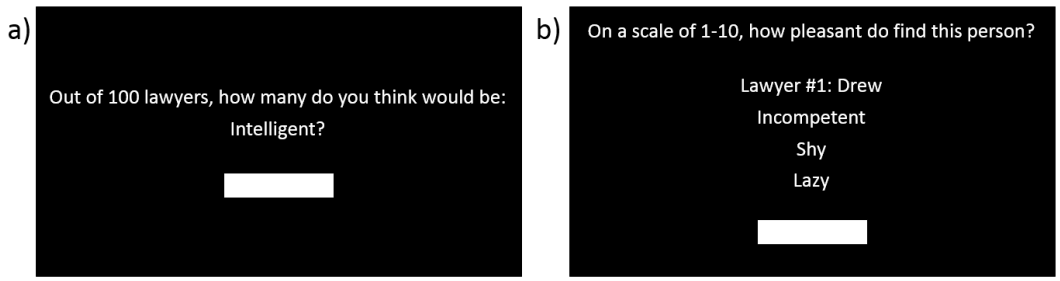

Figure 2.2: Sample slides from the measurement (a) and presentation (b) blocks of Experiment 1.

the category to measure any changes in expectation. This occurred after viewing 6, 18, 36, 60 and 90 total exemplars, with the ratio of traits within each interval being consistent with that of the complete exemplar set. At the start of each test block, participants were informed that though some of the questions had been asked before, they should answer based on how they felt at that point in time.

After viewing all 90 lawyer summaries and completing the final test block, the experiment ended, and participants were debriefed as to the aims and expectations of the study.

2.3.2 Results

Data Analysis

For ease of analysis, congruent and incongruent ratings were merged into a single score of stereotypicality to remove trait congruency as a factor; this was done by converting incongruent ratings by calculating the difference between each rating and the maximum score of 100 and averaging across the resulting ratings for each participant in each test block. Separate analyses for each trait type are available in the appendix. Neutral ratings, meanwhile, were excluded from analysis.

Figure 2.3 shows mean trait ratings from Experiment 1. The results of the experiment were analysed using a Bayesian repeated measures ANOVA including the factors of test block and concentration condition. A Bayesian ANOVA provides both the standard F statistics for each factor as well as a Bayes factor value BF_{inc}

measuring the relative evidence for the inclusion of that factor within the model against a null hypothesis of excluding that factor. To aid interpretation, BF_{inc} values of 3, 10 and 100 are respectively considered substantial, strong and decisive evidence for the alternative hypothesis, while values of 1/3, 1/10 and 1/100 are respectively considered substantial, strong and decisive evidence for the null hypothesis (Rouder, Speckman, Sun, Morey, & Iverson, 2009). All Bayes factors were calculated using the R package BayesFactor with a JZS prior as defined by Rouder et al. (2009) under the default prior scale of 0.707. As the first test block was intended to provide a baseline, being unaffected by either exposure to the exemplar set or concentration condition, ratings from this round were not included in the ANOVA; this assumption is assessed in the follow-up tests below.

The Bayesian repeated measures ANOVA found a significant effect of test block, $F(1,4) = 39.4$, $p < .001$, $BF_{inc} > 10000$, with ratings becoming less stereotypical over the course of the experiment. Concentration was, however, found to be non-significant, $F(1,4) = 0.99$, $p = .322$, $BF_{inc} = 0.33$, suggesting no difference in ratings between the two conditions. In addition, no significant interaction was found between test block and concentration, $F(1,4) = 1.83$, $p = .122$, $BF_{inc} = 0.19$, indicating the effect of exposure to the exemplar set also did not differ between the

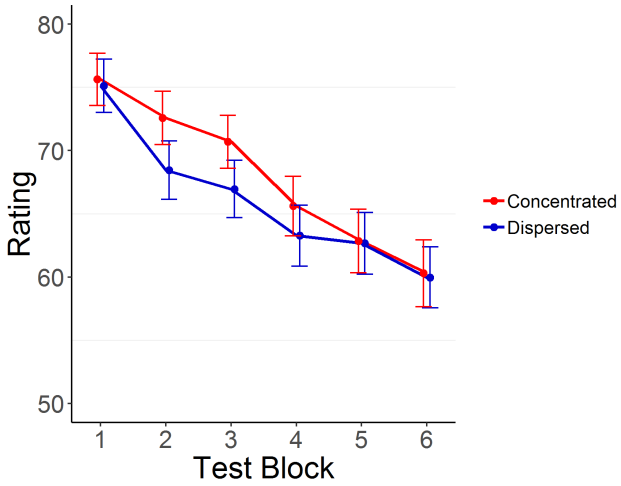

Figure 2.3: Mean trait ratings from the two concentration conditions across the six test blocks from Experiment 1. Error bars show 95% CIs.

two groups.

To further examine the time course of these results, a series of Bayesian independent t-tests were performed comparing ratings between concentration conditions in each test block. As with the Bayesian ANOVA above, Bayesian t-tests provide a Bayes factor BF_{10} for each comparison, here measuring the relative evidence for a given alternative hypothesis against the null hypothesis that the true difference in means is zero; these values can be interpreted using the same scale given above. Based on expectations from previous uses of the concentration design, these tests were one-tailed, with the alternative hypothesis suggesting ratings to be higher in the concentrated condition. This did not however apply to ratings from the first test block as this was intended to be a baseline unaffected by concentration; a two-tailed hypothesis was therefore used in this block to assess the accuracy of this assumption.

Results from these tests are summarised in Table 2.3; ratings were found to be significantly higher in the concentrated condition than the dispersed condition in the second and third blocks, whereas later blocks showed no significant differences between conditions. Bayes factors do however show that these differences do not reach the level of substantial evidence in favour of the alternative hypothesis, though this hypothesis is more likely than the null. The data then provides an indication of a concentration effect in early blocks which fades with further exposure, though this does not appear to be a substantial effect, as reflected in the interaction term above.

Block	t	df	p	BF_{10}
1	0.32	114	0.750	0.21
2	2.11	114	0.019	2.79
3	1.68	114	0.048	1.32
4	0.98	114	0.166	0.50
5	0.06	114	0.475	0.21
6	0.13	114	0.449	0.22

Table 2.3: Bayesian t-test results from Experiment 1

2.3.3 Discussion

The results of Experiment 1 are somewhat surprising: while ratings did become less stereotypical with exposure to the exemplar set, reflecting the high rate of counter-stereotypical data in these group members, the apparent lack of difference between concentration conditions suggests expectations were seemingly unaffected by the patterns in this data, thereby failing to replicate the subtyping effect of Weber and Crocker (1983). This also means that the collected data displays no change in subtyping with further exposure, a key prediction of three of the candidate models described above: the SSM predicts an increase in subtyping with exposure, the RMC predicts a convergence between conditions, while the RRMC predicts a reversal of subtyping at higher volumes of incongruent evidence where the subtype becomes the most likely cluster. The results therefore best correspond with the predictions of the BKM, with expectations being determined according to the prevalence of these traits in observed exemplars irrespective of the patterns shown in these category members.

There are, however, two issues which could be raised with this conclusion: first, follow-up tests comparing ratings in individual test blocks do show significant differences in early but not late blocks, suggesting the presence of a subtyping effect which fades with further exposure; while Bayes factors find these differences do not meet the criterion for compelling evidence of a subtyping effect in these blocks, this is a notable discrepancy with the predictions of the BKM, and could suggest more complex categorisation processes in behaviour than are offered by this model. Second, such a finding conflicts with previous displays of subtyping in past research, including studies using similar concentration designs to the present experiment (Weber & Crocker, 1983; Johnston & Hewstone, 1992; Bott & Murphy, 2007), suggesting that the predictions of the BKM are not accurate in all cases of stereotype change. As such, while the BKM does appear to be more likely to underlie the observed behaviour than the other candidate models, this requires further evidence before this can be accepted as an accurate depiction of the processes

involved in stereotype revision.

We therefore sought to further examine the processes underlying behaviour in this task using direct quantitative comparisons with simulated data from the candidate models. Such a comparison does however raise an ambiguity in the design of the task regarding exemplar representation which may interfere with model fitting: it is unclear how stereotype-neutral traits should be structured within the models. In the above experiment, neutral traits appeared on the same dimensions as stereotype-congruent and -incongruent traits following the exemplar structure of Weber and Crocker (1983). This is notable given that the neutral traits are supposedly distinct from stereotypical beliefs, and therefore should not have such a relationship; as such, these traits could be considered to sit on separate dimensions, providing a contrasting exemplar structure, as illustrated in Table 2.4. It is however unclear which of these formats was used by participants in this experiment: common dimensions may better suit the presentation of these exemplars, while separate dimensions may better suit the underlying logic of these traits. This presents a problem when defining this structure for the candidate models, as there is less assurance of a match between participant and model assumptions.

To address this ambiguity, we therefore performed a second experiment in which neutral traits were entirely removed from exemplars, thereby circumventing these potential issues; this then provided a second data set where participant estimates are definitively unaffected by these traits regardless of their format. Such a change to the exemplar structure does admittedly introduce potential differences

	Inferred Structure								
Exemplar	Common Dimensions			Separate Dimensions					
1	I	N	I	I	-	I	-	N	-
2	C	I	I	C	I	I	-	-	-
3	I	I	N	I	I	-	-	-	N

Table 2.4: Potential inferred exemplar structures from Experiment 1, where neutral values (N) could either be assumed to fall on the same dimensions as congruent (C) and incongruent (I) values, or placed on separate dimensions of their own, with '-' indicating a missing value.

in the resulting partition, though continued adherence to the concentration design should maintain focus on subtyping formats, contrasting scenarios where incongruent data is and is not easily isolated. In addition, this provided a partial replication of the first experiment, thereby offering further verification of the above suggestion that expectations are insensitive to the presentation format of counter-stereotypical information.

2.4 Experiment 2

Experiment 2 replicated the design and procedure of Experiment 1 with one key alteration: exemplars did not contain any stereotype-neutral trait values on any dimension, eliminating the ambiguity as to how these traits were treated by participants in Experiment 1.

2.4.1 Method

Participants

Ninety-nine participants were selected from the University of Warwick online recruitment system in return for £3 in payment. The sample included 61 females and 38 males, while age ranged between 18 and 41 years, with a mean of 22.8.

Design and Materials

Experiment 2 used the same basic design as Experiment 1, extending the concentration design of Weber and Crocker (1983) across a larger exemplar set and taking multiple measures of stereotypicality at set presentation intervals. The key difference, however, was in the structure of the two exemplar sets, with neither the concentrated or dispersed set containing any stereotype-neutral traits. This was done by editing the previous exemplars sets to replace all neutral values with congruent values; trait ratios in the new sets were therefore two-thirds incongruent and one-third congruent. Exemplar structure in this task is illustrated in Table 2.5. Exemplars

	Condition					
Exemplar	Concentrated			Dispersed		
1	I	I	I	I	C	I
2	I	I	I	C	I	I
3	I	I	I	I	I	C
4	I	I	I	I	C	I
5	C	C	C	I	I	C
6	C	C	C	C	I	I

Table 2.5: Exemplar structure from Experiment 2, replicating the concentration design of Experiment 1 without the use of stereotype-neutral traits.

were again said to come from the category of lawyers, so using the same congruent and incongruent trait labels, while the same set of names was used for individuation between exemplars.

Procedure

Experiment 2 used an identical procedure to Experiment 1 with one exception: participants were no longer asked to provide estimates of the appearance rate of the stereotype-neutral traits given their exclusion from exemplar summaries. As such, each of the six rating blocks asked only for ratings of the five congruent and five incongruent traits. Participants were again randomly assigned to one of the two concentration conditions, with 50 participants viewing the concentrated exemplar set and 49 viewing the dispersed exemplar set.

2.4.2 Results

Data Analysis

Figure 2.4 shows mean ratings from Experiment 2. Results from the task were analysed using the same procedure as Experiment 1, aggregating trait ratings into a single stereotypicality score and comparing this measure using a Bayesian repeated measures ANOVA using the factors of test block and concentration condition, again excluding ratings from the first test block. This found a significant effect of test block, $F(1,4) = 27.5$, $p < .001$, $BF_{inc} > 10000$, with ratings again becoming less

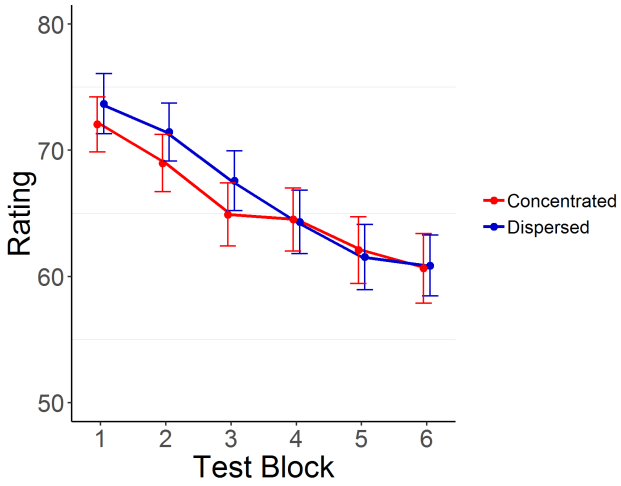

Figure 2.4: Mean trait ratings from the two concentration conditions across the six test blocks from Experiment 2. Error bars show 95% CIs.

stereotypical over the course of the experiment, but no significant effect of concentration condition, $F(1,4) = 0.11$, $p = .747$, $BF_{inc} = 0.27$, with ratings being no different between the conditions. Similarly, no significant interaction was found between test block and condition, $F(1,4) = 1.17$, $p = .322$, $BF_{inc} = 0.07$, indicating the effect of exposure to the exemplar set did not differ by concentration.

These results were again further examined using a series of Bayesian t-tests to compare ratings between conditions separately in each test block. As with the previous experiment, these tests were one-tailed based on the results of previous uses of the concentration design, using the alternative hypothesis that ratings would be higher in the concentrated condition; again however, this did not apply to the first test block, which was expected to be unaffected by concentration condition.

Bayesian t-test results from the second experiment are summarised in Table

Block	t	df	p	BF_{10}
1	0.79	97	0.434	0.28
2	0.93	97	0.823	0.12
3	0.92	97	0.821	0.12
4	0.06	97	0.476	0.22
5	0.19	97	0.426	0.25
6	0.03	97	0.513	0.21

Table 2.6: Bayesian t-test results from Experiment 2.

2.6; no significant differences were found between concentration conditions in any of the test blocks, with Bayes factors instead showing substantial evidence in favour of the null hypothesis in all cases.

2.4.3 Discussion

In contrast with the slight ambiguity of the findings of Experiment 1, the results of Experiment 2 are much clearer on the impact of the concentration manipulation: while ratings again demonstrated no difference between conditions overall, follow-up tests now support this result in all cases, offering substantial evidence that subtyping effects were not displayed in any test block. This also means that the data demonstrates no change in subtyping across the task, as was indicated in the first experiment, with the conditions remaining consistently similar throughout exposure.

Removing the neutral values from the exemplar descriptions therefore appears to have actually assisted in clarifying the results of the previous experiment, eliminating any evidence of a subtyping effect in the task entirely. The data from this task then provides a closer match to the predictions of the BKM, demonstrating no effect of the format in which counter-stereotypical information is presented. This is an interesting contrast with the first experiment, which, while not conclusive, did demonstrate some minor deviations from this pattern in early test blocks; it is unclear why such a change in exemplar structure should lead to less variations between conditions if the underlying categorisation process remains consistent. This could then indicate a degree of flexibility in the system producing these estimates, demonstrating some reaction to the differences in exemplar pattern between the two experiments. While such a reaction is not possible within the framework of the BKM, this does match with the flexibility of the previously noted clustering methods, adapting the representation of the category to suit observed data patterns rather than being restricted to a single heuristic rule. The behaviour observed in this task could then be attributable to a clustering mechanism which is emulating the process

of the BKM, but with some minor deviations according to variations in exemplar structure. This could then also explain why subtyping effects have been observed in previous studies but not in the current experiments, as differences in presentation format or experiment structure could generate differences in participant reactions. This is, however, a purely speculative explanation, as the collected results remain most indicative of a book-keeping process within the present tasks.

This does however mean that Experiment 2 is also unable to provide any substantial assistance in determining the rationality of subtyping; by trying to align participant and model assumptions, the changes to the design instead in fact seemingly further distanced the results from previous displays of subtyping, preventing any further insights into the process. It therefore appears necessary to maintain the structure of the concentration design as much as possible when examining this effect, meaning any adjustments to the neutral traits must be more delicate. As such, the third experiment in this section attempted to redefine the dimensions of stereotypicality by changing the labels presented on these dimensions rather than the underlying structure; this used a shift from the polarised personality dimensions of the previous experiments to discrete behavioural and physical traits, providing aspects that could be stereotypical, counter-stereotypical or unrelated to the stereotype whilst still explicitly existing on the same dimension. In addition, as this requires the generation of a new set of trait labels, this also provides an opportunity to extend the findings of the previous experiments to a novel group, allowing for greater assurance of the generalisation of the observed effects; the following experiment therefore replaced the lawyer category used in the previous tasks with the category of police officers, thereby targeting a different set of participant beliefs.

2.5 Experiment 3

Experiment 3 again replicated the design and procedure of Experiment 1 with some minor alterations, though in this case the changes to the task were mainly superficial, editing the labels attached to the trait values used in exemplar summaries rather

than the values themselves. This change in labels was intended to accomplish three main goals: first, to explicitly place neutral traits on the same dimension as stereotypical and counter-stereotypical traits for a more valid exemplar structure; second, to move from trait labels sitting at the ends of a continuum to discrete, categorical labels; and third, to assess the generalisation of the effects seen in the previous experiments to an alternative category with differing stereotypical beliefs.

2.5.1 Method

Participants

One-hundred-and-twenty-two participants were selected from a University of Warwick undergraduate psychology class as part of a course requirement. The sample included 98 females and 24 males, while age ranged between 18 and 25 years, with a mean of 18.7.

Design and Materials

As in Experiments 1 and 2, Experiment 3 used an extended form of the concentration design of Weber and Crocker (1983) using multiple measurement blocks. This used identical concentrated and dispersed exemplar sets to those of Experiment 1, but a different set of trait labels to represent these values, employing discrete behavioural labels in place of the more continuous personality labels used in the previous experiments. As this required the generation of a new set of trait labels for the task, a change was also made to the target exemplar category in order to assess the generalisation of previous effects to novel groups; exemplars were therefore said to belong to the category of police officers, chosen due to a high likelihood of both participant familiarity and existing assumptions.

Five new stereotype-relevant dimensions were therefore generated for the category of police officers through discussion between the authors, with multiple potential congruent, incongruent and neutral labels. In order to determine which labels were most reliably stereotypical, expectancies for these traits were tested in

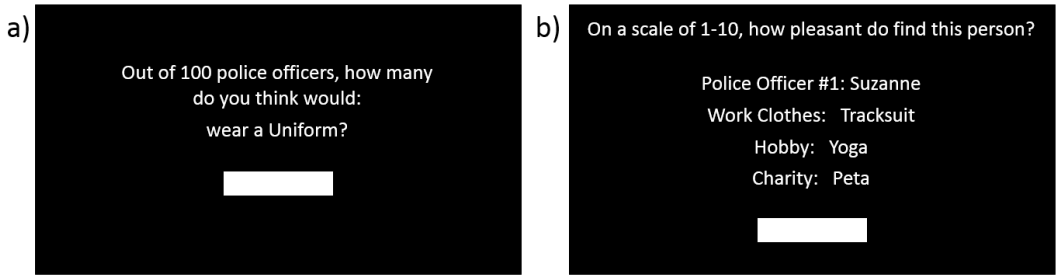

Figure 2.5: Sample slides from the measurement (a) and presentation (b) blocks of Experiment 3.

a pilot study (N=6). This used the same trait rating paradigm from a single test block of the previous experiments, here applied to 20 newly generated trait labels. Mean trait ratings were then used to classify the new labels as either congruent, incongruent or neutral. Trait labels were defined as congruent if mean ratings were above 60 and incongruent if below 40, with intervening values being considered neutral. Where multiple traits on a given dimension shared the same classification, the more extreme rating was selected to represent a congruent and incongruent values, while the trait with the lowest variance was selected for neutral values. This criteria eliminated two dimensions for failing to provide a congruent, incongruent and neutral trait label, leaving three dimensions for use in exemplar summaries: work clothing (congruent: uniform, incongruent: tracksuit, neutral: suit), hobby (congruent: football, incongruent: yoga, neutral: rugby) and supported charity (congruent: UNICEF, incongruent: PETA, neutral: Greenpeace).

Exemplars again also included a randomly selected name for individuation, taken from the same set used in the previous experiments.

Procedure

The procedure of Experiment 3 was identical to that of Experiment 1, replacing only the labels used in exemplar summaries and trait ratings to those given above for the category of police officers. Sample slides from this task are shown in Figure 2.5. Assignment to concentration condition was again randomised, with 61 participants viewing the concentrated exemplar set and 61 viewing the dispersed exemplar set.

2.5.2 Results

Figure 2.6 shows mean trait ratings from Experiment 3. Data was analysed using the same procedure as the previous experiments, using a Bayesian repeated measures ANOVA including test block and concentration condition, again excluding ratings from the first test block. As with the previous experiments, this found a significant effect of test block, $F(1,4) = 37.4$, $p < .001$, $BF_{inc} > 10000$, with ratings becoming less stereotypical over the course of the task, but no significant difference in ratings between concentration conditions, $F(1,4) = 0.40$, $p = .528$, $BF_{inc} = 0.23$, and no significant interaction between these factors, $F(1,4) = 0.27$, $p = .901$, $BF_{inc} = 0.01$.

This was again followed by a series of Bayesian t-tests between conditions in each test block to further examine these results. These tests were again one-tailed, predicting ratings to be higher in the concentrated condition, with the exception of the first test block. Bayesian t-test results for Experiment 3 are summarised in Table 2.7; as with Experiment 2, no significant differences were found between conditions in any of the test blocks, with Bayes factors again providing substantial evidence for the null hypothesis in all cases.

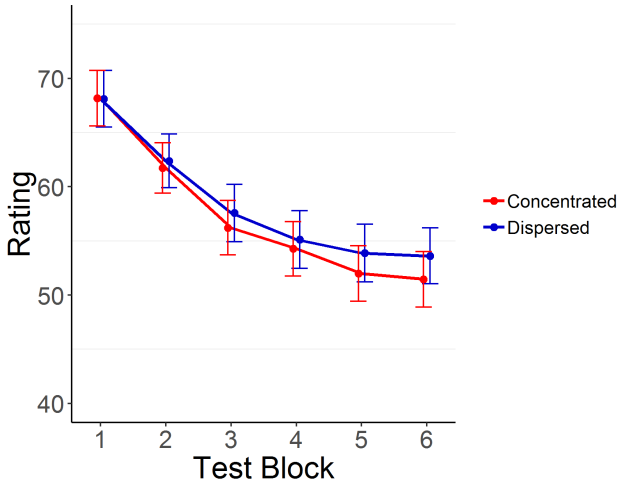

Figure 2.6: Mean trait ratings from the two concentration conditions across the six test blocks from Experiment 3. Error bars show 95% CIs.

Block	t	df	p	BF_{10}
1	0.03	120	0.974	0.19
2	0.33	120	0.630	0.20
3	0.57	120	0.716	0.22
4	0.33	120	0.628	0.20
5	0.70	120	0.757	0.24
6	0.83	120	0.795	0.26

Table 2.7: Bayesian t-test results from Experiment 3.

2.5.3 Discussion

The results of Experiment 3 correspond with those of Experiment 2: ratings did not appear to differ between concentration conditions at any point during the task, again indicating no subtyping effects in any test block and so no change in subtyping with further exposure. This again places the data most in line with the predictions of the BKM, with responses seemingly being unaffected by the organisation of new information. The data therefore again provides no further insight into the rationality of the subtyping effect discussed in the introduction to this study, instead suggesting a stereotype revision process in which subtyping is not possible.

The collected empirical data therefore shows a reasonably consistent pattern of evidence across the three tasks: all three experiments fail to reliably demonstrate the concentration effect of Weber and Crocker (1983), instead showing no difference in expectations according to patterns in exemplar data. This then supports a book-keeping model of stereotype revision in which beliefs are updated to reflect each new piece of evidence as it is encountered. This does however present a conflict with past demonstrations of subtyping effects in previous studies of stereotype change (Weber & Crocker, 1983; Johnston & Hewstone, 1992; Bott & Murphy, 2007), potentially leading to some concerns as to the generality of these results: book-keeping does not appear to provide an adequate account of the range of behaviours observed beyond the current study. This is in addition to the minor deviations from this pattern of evidence shown in Experiment 1 which could call these results into question. As such, in order to further explore these results, the collected data was next contrasted with simulated responses from the four candidate models

to determine which offers the most accurate depiction of behaviour. This provides a more direct quantitative assessment of the accuracy of these models, supplementing the above theoretical contrast.

2.6 Model Comparison

To generate direct, quantitative model predictions, the four candidate models were run through the same exemplar data presented to participants in each of the three experiments, taking equivalent measures of the probability of stereotypical traits appearing in the category at the same intervals, and comparing these predictions with the collected data. This used a grid point search function across model parameters to determine the best fit of each model to participant data, calculating the fit at certain pre-set combinations of parameter values to suggest the closest match to behaviour. A grid search was used due to potential issues with traditional gradient descent optimisation functions in clustering methods, which can have difficulty in navigating the complex likelihood function generated by such models. Grid points therefore varied the α parameter shared by all four models, as well as the coupling parameter c used by the RMC and RRMC, and the incongruency criterion parameter θ used by the SSM. Considered values for these parameters were: for α, 0.01, 0.1, 0.2, 0.5, 1, 2, 5, 10, 15, 20, and 30; for c, 0.01 to 0.99 in steps of 0.01; and for θ, 0.1 to 1 in steps of 0.1.

In addition to these parameters, as the category in question is a familiar social group with which participants are likely to have previous experience, all four models also included a set of exemplars added to the partition before exposure to the experimental exemplar sets in order to simulate such prior knowledge. These prior exemplars were used both to provide a more valid depiction of the origins of the group stereotype according to the ratio of congruency in this set, as well as to allow for potential interactions between prior knowledge and new information, as have been observed in other categorical modelling studies (e.g. Heit, Briggs, & Bott, 2004).

Each model therefore began by generating this set, with the number of prior exemplars n_p and the rate of congruency in the set p_c being included as additional parameters: this was done by randomly distributing congruent and incongruent values across the prior exemplar set, with the number of these values being defined by the number of prior exemplars, the number of stereotype-relevant dimensions and the prior rate of congruency. Exemplars in this set also included an additional dimension noting their membership of the target category. Once generated, these exemplars were then assigned using the methodology of the given model, emulating the partitioning of previously encountered category members. These parameters were therefore also included in the grid point search, with the considered values being: for n_p, 0 to 100 in steps of 5; and for p_c, 0.1 to 1 in steps of 0.1. As such, the BKM was defined as having three free parameters, while the remaining models had four.

The four models were run through the same exemplar sets given to participants at each combination of parameter values to generate estimates of the probability of both congruency and incongruency in new category members at each of the six exemplar intervals. For greater reliability in model estimates, the models were run 20 times at each grid point, and predicted probabilities were averaged across these trials. These mean values were then used to calculate model likelihoods assuming identical parameter values for all participants in order to allow the model to fit the two concentration conditions simultaneously.

Likelihoods were calculated at each grid point using the deviation between mean trait ratings and mean model predictions for each trait type in each test block, allowing all responses to be placed on a single distribution in order to fit all blocks and conditions simultaneously; this produced a set of 24 deviation values, representing mean ratings from the two trait types in the six test blocks from the two concentration conditions. These deviation figures were then converted into probabilities according to a normal distribution across deviation with mean 0 and variance fit to maximise the final likelihood product, providing a single likelihood value for each model at each grid point. Likelihoods were also aggregated across Experi-

ments 1 and 2 assuming a common set of parameters for those tasks given their use of the same target category and feature set. Experiment 3 was however fit separately as this used a distinct category, leading to a separate set of likelihood scores for this task. This then provided two maximum likelihood values for each model: one across Experiments 1 and 2, and a separate value for Experiment 3.

Maximum likelihoods from each model were then used to calculate Bayesian information criterion (BIC, Schwarz, 1978) values for comparison to account for differences in complexity between models; BIC figures provide an adjusted measure of model fit, with lower values indicating a better match to data. BIC values were then summed across experiments to provide a general measure of model fit across all collected data. These values were also used to calculate BIC weights to provide an estimate of the posterior probability of each model (Wagenmakers & Farrell, 2004).

This procedure also allowed for an investigation of the previously noted concerns regarding the format of stereotype-neutral traits in Experiment 1: to determine the assumed format of these traits, the models were also given alternate exemplar sets for partitioning in which neutral traits were moved to separate dimensions, allowing both formats to be compared with behaviour in this task. BIC values were found to be lower for the separate neutral format, indicating a better match to participant expectations, as shown in Table 2.8; as such, the following aggregated results are based on the use of this structure in that task. Such comparisons were not performed for Experiments 2 and 3 however given that neutral traits were removed from Experiment 2 and edited to definitively fall on stereotype-relevant dimensions in Experiment 3, removing this issue from these tasks.

Aggregate BIC scores across the three experiments are shown in Table 2.9.

Format	RMC	RRMC	BKM	SSM
Common Dimensions	140.57	142.03	142.47	169.05
Separate Dimensions	124.33	123.61	125.83	166.09

Table 2.8: BIC values for four candidate models across the two considered formats for neutral traits in Experiment 1.

Model	Parameters	MLL	BIC	w(BIC)
RMC	4	-176.04	380.27	0.878
RRMC	4	-178.08	384.35	0.115
BKM	3	-184.32	389.79	0.008
SSM	4	-236.43	501.06	0

Table 2.9: Aggregated modelling results from the three experiments, where MLL is the summed maximum log likelihood for that model across experiments, BIC values are summed across experiments, and w(BIC) is the weight of the BIC score when comparing the four models.

The RMC was found to have the best fit to the collected data, followed in order by the RRMC, BKM and SSM. BIC weights also show this to be a substantial advantage, as shown by the relative scale of these figures. We next discuss the predictions from these best fits for each experiment to assess the qualitative fit of the models.

2.6.1 Experiment 1

Predictions from the best fits of Experiment 1 are shown in Figure 2.7a. Best fitting parameters for this task were: for the RMC, $\alpha = 30$, $c = 0.22$, $n_p = 100$, $p_c = 0.9$; for the RRMC, $\alpha = 30$, $c = 0.21$, $n_p = 90$, $p_c = 0.9$; for the BKM, $\alpha = 30$, $n_p = 90$, $p_c = 0.9$; and for the SSM, $\alpha = 30$, $n_p = 100$, $p_c = 0.8$, $\theta = 1$.

When the predictions for this best fit for the RMC are examined, differences in probability estimates between concentration conditions for both measures are reasonably small, and appear to remain reasonably consistent across test blocks, contrasting with the apparent convergence between conditions observed in this task. This deviation is, however, put in context by the best fits of the competing models, which show greater differences between predictions and behaviour: the RRMC and SSM both show a greater divergence in later blocks, while the BKM by design predicts no difference between conditions, as well as a greater reduction in the strength of stereotypical beliefs than that seen in either the alternative models or the participant data. This reinforces that this comparison reveals only the best fit of the four candidate models rather than an absolute description of behaviour in the

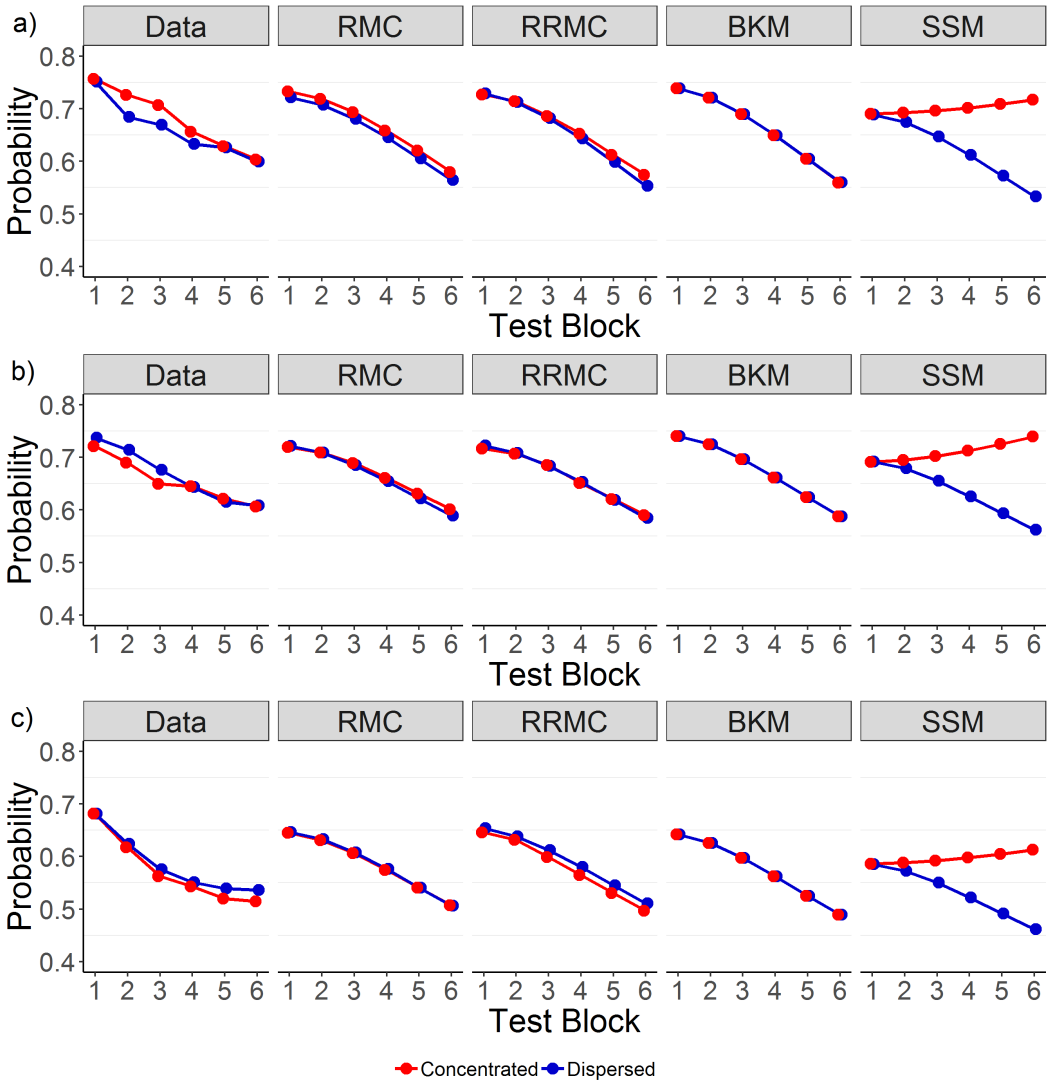

Figure 2.7: Trait probability estimates from the best fits of the four candidate models to Experiment 1 (a), Experiment 2 (b) and Experiment 3 (c), including empirical data adjusted to fall on the same scale for comparison.

task; more complex models may therefore be needed to reflect the subtle differences observed in the participant data.

2.6.2 Experiment 2

Predictions from the best fits of Experiment 2 are shown in Figure 2.7b. As noted above, best fitting parameters for this task were identical to those of Experiment 1.

Again examining these predictions more closely, both the RMC and RRMC

capture the observed data patterns reasonably well, mirroring the minor differences between concentration conditions, though these effects appear less substantial in the RRMC. The BKM, meanwhile, again naturally predicts no difference between conditions, and a greater fall in the stereotypicality of both measures than was observed in the experiment. Finally, the SSM again deviates substantially from behaviour, showing an increase in subtyping over the task, with measures from the two conditions moving in opposing directions.

2.6.3 Experiment 3

Predictions from the best fits of Experiment 2 are shown in Figure 2.7c. Best fitting parameters for this task were: for the RMC, $\alpha = 30$, $c = 0.24$, $n_p = 80$, $p_c = 0.8$; for the RRMC, $\alpha = 15$, $c = 0.18$, $n_p = 100$, $p_c = 0.7$; for the BKM, $\alpha = 20$, $n_p = 100$, $p_c = 0.7$; and for the SSM, $\alpha = 30$, $n_p = 100$, $p_c = 0.6$, $\theta = 0.7$.

The predictions of these best fits are fairly similar to those described above: the RMC shows little difference between concentration conditions, whereas the RRMC and SSM predict a larger difference which appears to grow over the course of the task, while the BKM again predicts no difference between conditions at any point and a greater reduction in stereotypicality than was observed in participant data.

2.7 General Discussion

In this chapter, we attempted to assess the underlying mechanisms of stereotype revision, examining responses to counter-stereotypical information both empirically and computationally in three different scenarios. This revealed a reasonably surprising set of results: across the three tasks, the present data offers no replication of previously observed subtyping effects, instead indicating a book-keeping revision process in which subtyping is in fact impossible. Such findings present a notable conflict with previous studies of stereotype revision (Weber & Crocker, 1983; Johnston & Hewstone, 1992; Bott & Murphy, 2007), making the determination of

the systems generating these effects difficult, with behaviour seemingly changing according to differences in experimental design. Model comparisons do however help to reconcile this conflict, providing a quantitative indication that participant responses were more likely to be generated by a flexible categorisation mechanism which adapts the representation to suit patterns in observed data rather than a pure book-keeping process. The present data may not then in fact provide evidence against past demonstrations of subtyping, but could instead present a different potential outcome of a common underlying system which may produce subtyping effects under different circumstances.

These results therefore provide two main conclusions regarding the processes underlying stereotype revision, each reflecting a distinction which can be made in the definition of these systems: parametric versus non-parametric processes, and rational versus irrational processes.

2.7.1 Parametric vs. Non-parametric Processes

The four models presented in this study can be divided into two classes: the more basic parametric systems offered by the BKM and SSM which provide heuristic rules for stereotype revision, and the more advanced non-parametric clustering systems of the RMC and RRMC which use a more flexible representation. Within this distinction, the data collected here provides a case for non-parametric over parametric systems, demonstrated in the fits of the candidate models to behaviour: the RMC and RRMC substantially outperform the BKM and SSM, suggesting a strong sensitivity to the specific scenario in which category information is received. Such a finding corresponds with the adaptable nature of the non-parametric systems in which the generated representation reflects both new data patterns as well as potential interactions with pre-existing knowledge; this is in contrast with the less flexible parametric rules, which are unable to account for such variation in outcomes. As such, our findings indicate that the mechanisms underlying stereotype revision involve the incorporation of new information into a highly responsive men-

tal representation, with subsequent judgements being coloured by the combination of both old and new data patterns. Conversely, these findings argue against more heuristic depictions of revision which rely on static rules for the use of new data, as such systems are simply too restrictive to accurately capture the flexibility of human behaviour.

This also depicts the maintenance of stereotypical beliefs associated with subtyping as a natural aspect of such a categorisation process rather than a distinct strategy of stereotype preservation: subtyping occurs where a particular data pattern inadvertently diminishes the impact of counter-stereotypical data via isolation within a distinct subgroup, while other data patterns may generate different results. Such a concept is supported by theories of 'subgrouping' (Richards & Hewstone, 2001), a counterpoint to subtyping where the category is divided into lower-order groups without a separation between stereotypical and counter-stereotypical information. This has previously been seen to lead to greater levels of stereotype change (e.g. Maurer, Park, & Rothbart, 1995), contrasting with the maintenance of beliefs associated with subtyping: subgrouping is suggested to allow for a more even integration of stereotype-incongruent information rather than the isolation and exclusion associated with subtyping. Subgrouping may therefore represent an alternate outcome of the same categorisation process, creating a different partition which aids stereotype change; this then presents subtyping and subgrouping not as opposing strategies, but two potential outcomes of a common system. It should be noted, however, that the tasks used in this study provide no direct measure of the partitions used by participants, simply inferring this structure from observed predictions. It may then be advisable to include measures of partitioning in future work to determine whether such inferences correspond with reported groupings, for example using sorting tasks to provide a comparison with model predictions.

It is also notable that such a system presents subtyping as a reasonably fragile phenomenon, being displayed in only a small subset of scenarios which facilitate such isolation: indeed, the present study finds subtyping effects are difficult to effectively replicate, with none of the three experiments above finding any effect of

data pattern based on empirical data alone. This could then indicate that stereotype maintenance associated with subtyping is less of a cause for concern as it may initially seem, appearing only under fairly restrictive circumstances. This will however need to be assessed in more diverse settings to determine the precise scenarios in which subtyping occurs, including different target categories and exemplar formats beyond those used here.

This distinction also offers an interesting comparison of the three theories of stereotype change noted in the introduction to this section: book-keeping, conversion and subtyping. While the collected empirical data does appear to generally support book-keeping, model fitting shows that this theory is ultimately insufficient to capture the variations in behaviour displayed by participants in these tasks, being unable to reflect the apparent sensitivity to patterns in observed data. Simultaneously, these behaviours also do not strictly adhere to either pure subtyping or conversion, showing gradual changes in beliefs to reflect observed information; behaviour in these tasks then appears to fall between the three theories, demonstrating greater flexibility than is offered by any one of these models alone. This further demonstrates that stereotype revision does not appear to rely on a single universal strategy, but reacts to data structures as suggested by the considered non-parametric models; indeed, the two poorly performing parametric models in the above comparison were both designed to specifically reflect one of these three theories, with the BKM emulating book-keeping and the SSM emulating subtyping. This then implies that these three theories are not independent processes of stereotype revision in their own right, but three potential behaviours that could be generated by the true system; this would again fit with the use of clustering mechanisms where the representation is flexible and adapts to incoming data patterns, thereby potentially offering all three of these outcomes, as well as more intermediate behaviours.

2.7.2 Rational vs. Irrational Processes

The above distinction therefore indicates that stereotype revision is more likely to rely on the non-parametric processes offered by the RMC and RRMC than the parametric systems of the BKM and SSM. This second distinction however creates a division between these two non-parametric models according to their rationality, as captured by their use of exemplar data: the RMC includes data from all exemplars within the partition in its predictions, making it a rational system, whereas the RRMC focuses on only one cluster in its predictions while others are excluded, and as such could be labelled an irrational system. The contrast between these models then reflects the question that formed the basis of this study of whether learners do in fact ignore relevant category data when making judgements, as described in the initial depictions of subtyping; use of the RRMC would allow learners to display such an irrational behaviour without falling back on the more extreme and inflexible form of exclusion offered by the SSM.

The collected data is less conclusive across this distinction: without any display of subtyping in any the three experiments, the present empirical results are unable to provide any evidence as to the rationality of this effect, seemingly supporting neither the RMC nor the RRMC. Computational results do help to clarify these findings however, showing the RMC to have a better fit across all collected data; this does then indicate that participants made use of a rational process to provide their estimates in these tasks, using all available category data when making associated predictions.

The present findings therefore suggest that stereotypes are in fact based on a rational categorisation of social group data which both identifies latent patterns within the category and uses the sum of these patterns to make new predictions, as revealed through the response of these beliefs to new data. This then also depicts subtyping itself as a more rational effect than it may initially seem: incongruent information is not disregarded in these scenarios due to a biased strategy of stereotype preservation, but may be mitigated based on broader data structures. The remain-

ing ambiguity in the present results does however suggest that further comparisons are required to provide more confidence in this conclusion; this particularly applies to empirical contrasts which are better able to determine whether subtyped data is excluded from the category, an element which remains uncertain in the current experimental designs.

2.7.3 Additional Factors

To close this section, we also note some additional factors regarding stereotype use which may need to be considered in future investigations, beginning with the level of interaction with counter-stereotypical data. It is notable that the present study focused on responses to fairly minimal interactions with counter-stereotypical information: the effects observed in all three experiments result solely from the observation of member summaries rather than any significant interaction with actual counter-stereotypical group members. While this does seem to be effective in changing beliefs within these tasks, it remains unclear whether these changes will persist over a longer time period, particularly outside of the laboratory environment: past research has often suggested that meaningful change requires intensive, long-term interaction with out-group members to generate a genuine reduction in stereotypical beliefs, best exemplified by the Contact Hypothesis (Allport, 1954). Further testing may then be required to assess whether such low-level interactions truly offer an effective path to stereotype revision, for example using more long-term measurements to examine the retention of any generated change. Alternatively, similar tests could make use of more substantial interactions to investigate whether the current observations are greater for such exposure; while the representation of level of interaction within the current model remains uncertain, one simple method would be to treat more substantial interactions as multiple observations in the partition, essentially viewing that individual as providing more data than a single exemplar. This suggestion should, however, be pilot tested to determine the validity of this representation before being incorporated into the model.

It is also notable that this study infers the representation underlying stereotyping through the response to new information rather than the initial acquisition of these beliefs. The present research aimed to examine existing theories of stereotype change using corresponding categorisation models, thereby providing a window into the operation of stereotype revision. This does however presume that participants not only hold these beliefs prior to any experimental manipulation, but also base these beliefs on actual prior experience with the target category. This may be a problem if stereotypes are not in fact built on such experience, for example being taught by a third party; this could alter not only the base representation, but also the malleability of any associated beliefs. It may therefore be necessary to more closely examine the origin of stereotypes alongside their response to new information in order to provide a more complete picture of stereotype use.

2.7.4 Conclusion

The present study provides a starting point for a rational approach to stereotype use, providing both theoretical, empirical and computational evidence that a rational model of stereotype change, while not universally accurate, does provide a reasonable account of behaviour both in these experiments as well as previous studies into stereotype maintenance. We therefore hope that this study can act as a foundation for continued work in this field, allowing subsequent research to further refine the presented models to provide a more accurate depiction of behaviour. This will serve to provide greater clarity regarding the operations underlying stereotype maintenance, and so aid in finding more potential methods for encouraging stereotype change.

Chapter 3

The Role of Numerical Format in Estimation

When building a representation of the environment, the form of this representation is not just determined by external data patterns, but also the assumed structure of the environment prior to any actual observations. The nature of these assumptions can therefore have drastic impacts on behaviour, affecting not only the representation that is ultimately formed, but also the actions that are taken based on this representation. In this chapter, we examine a specific case of the use of such assumptions within numerical estimation, contrasting different prior numerical structures both through experimental manipulations and model comparisons. This uses differing reactions of these structures to uncertainty as a method of distinction, providing a window both into the underlying form of the representation as well as the potential behavioural outcomes of the use of such a structure.

3.1 Numerical Estimation

In many everyday tasks, we are required to make quick estimates of discrete stimuli based on noisy perceptual data: the number of people in a crowded room, or cars in a lane of traffic, for example. These decisions are not solely reliant on perceptual information, but also use past experiences with such stimuli to guide responses: if

estimating the number of people in a room, the actor may consider similar occasions where that number was later provided and use this information to inform their decision. Such guidance in fact becomes increasingly valuable at higher values as people's ability to discriminate between figures decreases (Krueger, 1984; Izard & Dehaene, 2008). Accurate estimates are therefore reliant on the learning of the distribution of such figures, building representations that reflect the prevalence of these values in the real world.

The influence of such previous experience is in turn however dependent on its representation, reflecting the different forms in which numerical information could be stored. Existing research has offered two potential forms for such information in two contrasting number systems, each suggesting distinct impacts on new decisions: the approximate number system and the symbolic number system. The approximate number system refers to the innate understanding of numerosity displayed by both humans and animals in which numbers are conceptualised in a continuous analogue form (Dehaene, 2011). Storing prior experiences in this format should therefore lead future estimates to focus on values similar to those previously seen; if the previous room contained 50 people, then nearby figures such as 49 or 51 would also become more likely (e.g. Gershman & Niv, 2013). In contrast, the symbolic number system is the discrete verbal format learned in later life which allows for more complex mathematical operations (Izard & Dehaene, 2008); in this case, only the experienced value would increase in expectancy, making that response alone more likely in subsequent estimates. Such a representation would allow the learner to acquire reasonably complex distributions through experience, tracking the individual appearance rate of each potential value (e.g. Sanborn & Beierholm, 2016). This would, however, also be possible using a sufficiently complex continuous format: narrow similarity functions could emulate discrete formats, making it difficult to distinguish between these forms.

This then raises the question of which of these systems underlies discrete estimates: symbolic representations could be used to suit the discrete nature of responses and feedback, while continuous forms may be used in spite of these el-

ements to suit the more analogue perceptual data and translated into discrete figures as required. Despite the impact of this distinction on both the representation formed and the resulting behaviour, this has received little attention in previous research. What is more, what work has been done has found conflicting results, with studies finding evidence for both continuous (Gershman & Niv, 2013) and discrete (Sanborn & Beierholm, 2016) underlying systems.

The current study therefore attempts to separate these forms using two complementary methodologies: first, an empirical contrast taking advantage of a difference in the definition of simplicity within continuous and discrete representations, and second, a quantitative contrast between computational models of behaviour in this task. In the following sections, we introduce potential models of estimation following such discrete and continuous formats, examine the principles of these models to derive methods of distinction, and use both empirical and computational comparisons to provide insight into the representations used in numeric estimates.

3.1.1 Using Prior Experience

We begin by examining the process by which past estimates could be used to inform new judgements. While this has not been studied extensively in estimation, one existing theory which touches on this process is calibration; in this theory, past trials are suggested to be used as anchoring points to map a discrete response scale onto continuous numerical representations to make subsequent estimates more accurate (Krueger, 1984; Izard & Dehaene, 2008). In this case, numerical data is automatically encoded in a continuous format and translated into discrete figures as required; for example, Izard and Dehaene (2008) suggest an affine transformation between continuous and discrete formats, using parameters to adjust both the shape and position of the discrete response scale. Calibration therefore increases the accuracy of this translation by tuning these parameters to suit the observed data, better mapping these two scales against one another to improve all future estimates. Such a transformation is, however, limited in the probability distributions it is able to rep-

resent; while this may be sufficient for reasonably simple structures, more complex distributions such as those with multiple modes cannot be accurately represented by this process. This stands in contrast to empirical data showing that learners can in fact acquire such multimodal distributions (Sanborn & Beierholm, 2016; Gershman & Niv, 2013). What is more, these studies also provide evidence against the use of a more complex translation function (Sanborn & Beierholm, 2016), thereby suggesting the learning of these forms is reliant on other mechanisms than calibration. More flexible systems are therefore required to accurately represent these more complex forms.

An alternative framework for the use of past experience is provided by Bayesian Decision Theory (BDT), in which prior assumptions regarding the distribution of the target stimuli are combined with direct observational data to form a posterior distribution from which a response can be selected; feedback from this response can then be used to update the representation for use in subsequent estimates. Previous observations are therefore used to inform new responses by constructing a mental representation of the true distribution, noting the prevalence of particular values. This provides BDT with an advantage over calibration as it can capture more complex learning structures such as the multimodal distributions noted above, with estimates reflecting both current perceptual data as well as the history of past observations. BDT may then provide a clear and established method well suited to the modelling of numerical estimation, better capturing the underlying process. In fact, BDT has been previously used as a description of the estimation process within continuous motor responses (Kording & Wolpert, 2004; Acerbi, Vijayakumar, & Wolpert, 2014; Chalk, Seitz, & Series, 2010), further supporting its use in the present study.

The use of BDT also facilitates the current comparison between discrete and continuous representations: while the general principles of BDT may remain fixed, the definitions of individual elements can vary, allowing for contrasts between alternate Bayesian models with different representational formats. Here, this applies primarily to the structure of the prior distribution, as this provides the assumed

model of the environment, and so the representation of numerical information. The current study therefore focuses on contrasts between differing definitions of the prior, while other model elements remain identical. What is more, BDT also allows for such a distinction without necessarily assuming that such mechanisms are in fact used by real learners, instead only providing useful descriptions of actual behaviour (Tauber, Navarro, Perfors, & Steyvers, 2017). As such, the models presented here are considered descriptive rather than normative, placing the focus on the use of discrete and continuous numerical formats rather than the optimality of behaviour.

Continuous prior formats are provided by a number of systems, though the present study focuses on mixtures of Gaussian components due to the flexibility of such a representation, allowing for emulation of other continuous distributions. In a Gaussian mixture, observations are grouped together based on similarity to form a set of subgroups, each described by a Gaussian distribution, which can then be combined into a single prior (Rosseel, 2002; Vanpaemel & Storms, 2008; Anderson, 1991); these have been previously used within Bayesian models of continuous estimation (e.g. Acerbi et al., 2014), providing some basis for their use as a continuous candidate in the present contrast. Such a prior holds the advantage of flexibility, being able to adjust the number of components used in the representation to best suit observed data patterns rather than using a predefined component structure. This flexibility has led to the application of Gaussian mixture priors to discrete estimates in spite of their continuous format; one demonstration of this is provided by Gershman and Niv (2013), in which a Gaussian mixture prior was used to model the merging of distinct categories of discrete stimuli where these categories shared similar statistical features. In this case, the Gaussian mixture is suggested to allow for simplifications of the final representation due to a prior preference for fewer components in the distribution; this could then indicate that discrete estimates may benefit from the use of a Gaussian mixture prior in terms of cognitive economy or greater generalisability. Both continuous and discrete estimates could then make use of a common underlying estimation system which is able to adapt to the needs of the task to provide the most valuable representation, considering both the accu-

racy and simplicity of the resulting form.

Discrete prior formats, conversely, are provided by distributions such as the categorical prior, which can be used to record the appearance rate of each observed value, relying more on memory for past observations than an inferred statistical distribution. Such a prior may be better suited to numeric estimates given its greater correspondence to the discrete nature of stimuli and responses; learners could then use this prior under the assumption that this structure is more appropriate to the nature of the task. This would, however, potentially lead to differences in behaviour according to the differing world models implicitly assumed by these prior structures. To illustrate, consider the above application of simplicity according to component count from Gershman and Niv (2013) to both the Gaussian mixture and categorical priors: in the case of the Gaussian mixture prior, a preference for fewer components is assumed to lead to the merging of subgroups where possible, leading to a smaller number of broader, more varied components. Categorical components, conversely, are discrete tallies of identical value observations and cannot be broadened in this way, meaning a reduction in the number of components would instead reduce the number of values considered in the distribution. The same fundamental principle therefore leads to widely different outcomes for these two structures, with the Gaussian mixture prior considering more values in its final posterior and the categorical prior considering fewer, demonstrating the impact of this representational format on actual estimations. To return to the previous example of counting people in a room, simplicity in the discrete case means restricting responses to a limited set of answers (e.g. low/medium/high or nearest 10), while in the continuous case, responses could focus on a single mean value, but with what could be substantial departures.

It is therefore necessary to examine the prior structures used in numerical estimation to determine whether this process relies on specialised discrete formats suiting the discrete nature of this task or more general continuous forms that can be shared with other stimuli. This has in fact been previously investigated in a study by Sanborn and Beierholm (2016) in which participants performed a dot numeration

task using an underlying bimodal distribution (illustrated in Figure 3.1a). In this task, participants were asked to estimate the number of dots appearing on-screen, with responses being followed by direct feedback noting the true dot count, making both participant responses and task feedback discrete and definitive, so providing clear evidence of a discrete task structure. Behaviour in the experiment was then compared with Bayesian models of estimation using differing definitions of individual model elements, including a contrast between continuous and discrete prior formats using a categorical prior and a kernel density estimate. Results from this study found behaviour was better described by the categorical prior than the kernel density estimate, suggesting that participants were using a discrete prior structure in line with the discrete nature of the task.

The findings of Sanborn and Beierholm (2016) therefore indicate that discrete estimation makes use of similarly discrete elements in order to assist in constructing more precise mental representations. There is one caveat to this finding, however: while model comparisons did suggest participant behaviour was most likely to be based on the use of discrete structures, this result could also be produced by a mixture of continuous components under certain circumstances. This is due to the previously noted flexibility of the Gaussian mixture prior: by grouping similar values together, the Gaussian mixture is able to adjust the variance of its components to suit the observed data, allowing for both broad, highly varied clusters and narrow, focussed clusters. Such narrow clusters could then essentially emulate the components of a categorical prior in which all members are identical, making the component variance zero. This concern is in fact raised in the third experiment of Sanborn and Beierholm (2016), noting that such a complex Gaussian mixture could capture the true categorical structures: a mixture prior using narrow components at the modes of the distribution and a broader component across the midrange offers a reasonable approximation of the true bimodal form (illustrated in Figure 3.1b). While that experiment did attempt to control for this possibility by using a quadrimodal distribution where such emulation is less precise, this only excluded a narrow set of mixture forms, while more complex structures are still

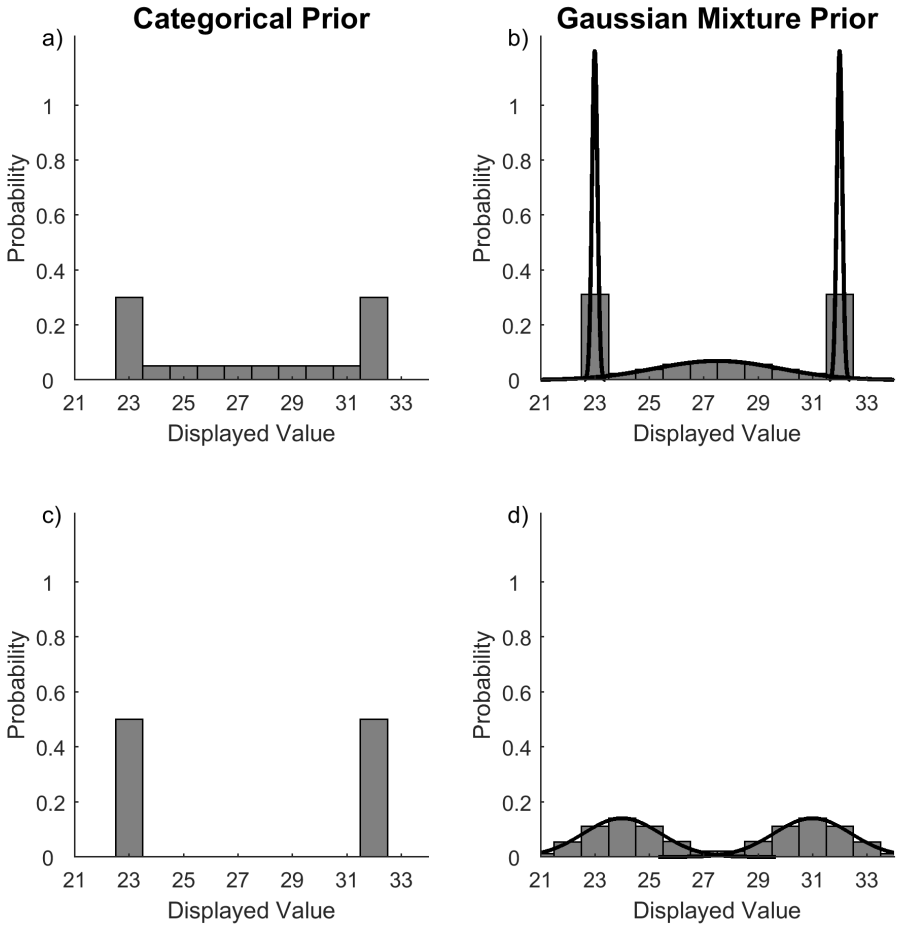

Figure 3.1: Comparison of the categorical (a) and Gaussian mixture (b) priors applied to the bimodal distribution of Sanborn and Beierholm (2016). Here, the categorical matches the true distribution, and the Gaussian mixture provides an approximation, with the black lines reflecting the individual distributions of each cluster. The lower figures demonstrate the proposed impact of uncertainty on the representation, leading to fewer potential response values in the categorical (c), but greater bleed-over in the Gaussian mixture (d).

possible. As such, the results of Sanborn and Beierholm (2016) can be explained in two different ways, with different implications: participants may have been using a more precise discrete prior in accordance with the discrete nature of the task, or a more flexible Gaussian mixture prior in line with that used for continuous estimates.

It is therefore necessary to distinguish between these explanations in order to determine whether discrete estimations do indeed rely on discrete structures, or

whether this was simply emulated by an otherwise continuous representation. As such, the present study aimed to perform a comparison between Bayesian estimation models using either a categorical or Gaussian mixture prior in a comparable estimation task; this builds on the results of Sanborn and Beierholm (2016) by examining a full continuous mixture model rather than one possible form of this prior for a more complete contrast of these formats.

While such a comparison provides a quantitative indication as to the underlying processes of numerical estimation, we also sought to supplement this contrast with a more qualitative investigation; this was intended to provide both a second method of distinction between prior formats as well as a demonstration of their opposing implications for actual behaviour. This distinction therefore drew on the previously noted differences between prior formats when applying principles of simplicity: while both priors are likely to prefer a lower number of components to simplify the final distribution, this takes two different forms according to the structure of these components, with the Gaussian mixture prior preferring to group more observations together to produce broader components, and the categorical prior limiting the number of values considered in the distribution to only a few key figures. It should then be possible to reveal which of these priors is used in this task by encouraging a reduction in components and observing which of these two reactions is displayed: the Gaussian mixture prior should move towards broader components, thereby covering more potential values and so allowing for more varied responses, while the categorical prior should focus on fewer potential responses, most likely limiting a bimodal such as that used in Sanborn and Beierholm (2016) to only the modes of the distribution, essentially turning the task into a high/low classification problem (illustrated in Figure 3.1c and d). This could be achieved by introducing uncertainty to the existing design of Sanborn and Beierholm (2016); if the true value of an observation is uncertain, both structures are likely to assign that observation to an existing component rather than assuming the presence of a new component.

It should then be possible to identify whether learners are using a truly discrete categorical prior or a continuous Gaussian mixture prior in this case by in-

troducing uncertainty to the dot numeration task of Sanborn and Beierholm (2016) and observing its effect on behaviour. The best method to achieve this is to cause doubt in the feedback given during the task whilst still providing the true value of the observation: if participants were made to distrust the feedback, for example by stating that this information was accurate in only a subset of trials, participants would no longer be able to rely on the definitive figures offered in the original design, likely leading to more confusion between actual values based on perceptual data. This allows for the addition of uncertainty to the task without changing any of the specific elements of the stimuli or feedback, instead changing the wider context of this information. What is more, such a manipulation represents a fairly valid scenario; real-world feedback is not always as reliable as that used in laboratory studies, potentially being noisy or vague, or originating from an untrustworthy source. In addition, this design also provides a simple method of manipulating the degree of uncertainty according to the apparent accuracy rate of feedback, allowing for an easy comparison between high and low levels of uncertainty.

The following experiment therefore sought to investigate the processes underlying numerical estimation by adding such a feedback uncertainty manipulation to a numerical judgement task in which participants were trained on a complex distribution through experience. This then provides a contrast of the competing hypotheses of the two potential formats introduced above: if participants are using a categorical prior, responses should be more polarised where feedback is less reliable, focussing mainly on the modes of the distribution. In contrast, if participants are using a Gaussian mixture prior, responses should be more spread out in this case, leading to more midrange and out-of-range responses. This also provided behavioural data for comparison with computational models of the task following these formats for a quantitative suggestion of the underlying process.

3.2 Experiment 1

3.2.1 Method

Participants

Forty University of Warwick students were recruited as participants in the experiment from the university's online SONA system in return for £8 in payment. The sample included twenty-five females and fifteen males, while age ranged between 18 and 39 years, with a mean of 22.4. While participants were paid for participation, these payments were not specifically tied to performance in the task.

Design

The experiment used an edited form of the dot estimation task of Sanborn and Beierholm (2016) in which participants were trained on an underlying distribution of dot values through an extensive series of estimation trials: in each trial, a number of dots appeared on the screen for 400 milliseconds, and participants were asked how many they believed had appeared. Dot counts were sampled from a bimodal distribution, ranging between 23 and 32 dots, with modes at the extremes of the range (illustrated in Figure 3.1a).

After giving each estimate, a feedback slide appeared noting both the participant's response as well as the actual dot count. In order to induce uncertainty in the feedback, the actual count was presented as a response given by a previous participant for that trial, with the level of uncertainty being manipulated according to the previous participant's reported accuracy rate across all estimation trials. The experiment therefore made use of a between-subjects uncertainty manipulation, using two uncertainty conditions: a high-uncertainty condition, in which the previous participant was stated to be accurate in 70% of trials, and a low-uncertainty condition, in which the accuracy rate was stated to be 95%. This rate was noted on every feedback slide to ensure participants were aware of uncertainty information.

A discrimination task was also used in the experiment to assess the partic-

ipant's discrimination ability for use as a parameter in later analysis. In the discrimination task, two sets of dots appeared sequentially on screen, and participants were asked which set (1 or 2) they believed to contain more dots. This was then followed by a feedback slide noting whether the response was correct or incorrect; this was not however affected by the uncertainty manipulation applied to feedback in the estimation task, being definitively accurate in all trials.

Procedure

Upon arriving at the lab, participants were first randomly assigned to one of the two uncertainty conditions, determining the reported rate of accuracy in feedback values. This was balanced to provide equal numbers of participants in each condition, meaning 20 participants were assigned to the high-uncertainty (70%) condition and 20 participants were assigned to the low-uncertainty (95%) condition.

Participants were told the experiment examined how decisions were made under uncertainty, and would involve estimating the number of dots appearing on screen. Participants first performed a set of 128 discrimination trials to assess their initial discrimination ability; this began with a series of 4 practice trials at low dot counts (1-4) to introduce the task.

After this first discrimination block was completed, participants then moved to the estimation task, again beginning with a set of 3 practice trials at low dot counts to introduce the task. Participants performed 500 total estimation trials, with breaks every 50 trials.

Once all estimation trials were completed, participants then performed another round of 128 discrimination trials to track any improvement in discrimination ability. Finally, participants were debriefed as to the aims and expectations of the study.

3.2.2 Results

Data from one participant was removed from analysis for failing to provide any responses within the presented dot range, leaving 39 subjects for comparison, with 19 in the 70% condition and 20 in the 95% condition. Responses further than 10 points outside of the displayed range were classified as response errors and removed from analysis; this eliminated an average of 1.81% ($\pm$ 0.46% 95% confidence intervals) of responses across participants.

Figure 3.2 shows conditional response distributions from the two uncertainty conditions, illustrating the average response rate for each presented dot value. Both groups demonstrated reasonable acquisition of the bimodal structure, showing strong preferences for the modes of the distribution in their responses; comparisons of each participants' mean number of responses across the two modes of the distribution with their mean number of responses across the eight remaining values found significantly higher numbers of modal responses, $t(38) = 5.33$, $p < .001$, $d = 1.42$. Unshown values were however also used as responses in both conditions, in keeping with the bleed-over predicted by the use of a continuous prior.

The key empirical contrasts from Experiment 1 are summarised in Table

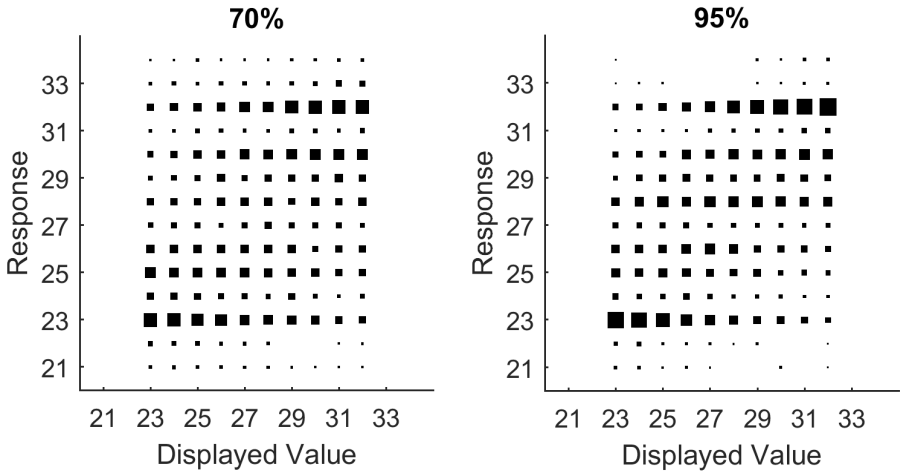

Figure 3.2: Conditional response distributions from the 70% and 95% uncertainty conditions of Experiment 1, where square size is proportional to the percentage of responses made to each displayed value.

Condition	70%	95%
Unique response count	17.4 (± 1.69)	15.3 (± 1.36)
Out-of-range responses	54.1 (± 22.3)	30.5 (± 23.8)
Mid-range responses	268 (± 55.9)	250 (± 46.1)

Table 3.1: Mean measures with 95% confidence intervals from the two uncertainty conditions of Experiment 1.

3.1. Analysis began by contrasting the count of unique responses from the two conditions: this was found to be significantly higher in the 70% group, $t(37) = 2.06$, $p = .047$, $d = 0.69$, with these participants using a wider range of values in their answers. No significant difference was found between the 70% and 95% groups however in either the number of responses from outside the dot range, $t(37) = 1.51$, $p = .140$, $d = 0.51$, or the number of mid-range (non-mode) responses, $t(37) = 0.54$, $p = .590$, $d = 0.18$, though both were found to be higher in the 70% condition.

The data therefore provides some support for the predictions of the continuous mixture prior: while participants in the high-uncertainty condition did not reliably offer a higher number of non-modal responses compared to the low-uncertainty condition, these participants did use a wider range of values in their responses, suggesting the use of a broader set of components when feedback was unreliable. This then provides limited evidence that numeric estimates rely on continuous numerical formats despite the discrete nature of stimuli and responses, utilising the inherent flexibility of such a system to adapt the representation to best capture external data patterns. The lack of reliable differences in all behavioural comparisons does however weaken this conclusion, meaning more substantial evidence is required before this suggestion can be accepted.

In order to address this concern and provide more confidence in the above conclusion, we decided to run a second experiment to further investigate this distinction using the same design but an alternate underlying distribution intended to provide a clearer separation between the two models. This followed the design of the third experiment of Sanborn and Beierholm (2016) in which a more complicated quadrimodal distribution (illustrated in Figure 3.3) was used in place of

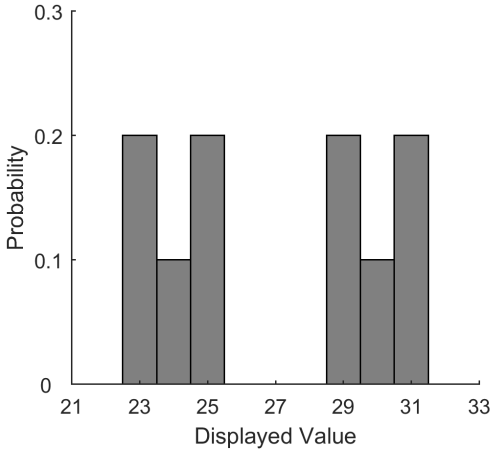

Figure 3.3: The quadrimodal distribution used in Experiment 2.

the initial bimodal as a method of further distinguishing between categorical and Gaussian mixture formats: such a distribution is more difficult to emulate using a mixture of continuous components, making the two prior formats more distinct. The use of such a distribution in the present study also provides a clearer separation in empirical measures: the quadrimodal provides a set of values in the middle of the displayed range that are not used in feedback, but may benefit from bleed-over from the two nearby modes under a continuous format. As such, if estimates in this task are in fact based on continuous prior structures, the use of a quadrimodal distribution should offer a clearer demonstration of these effects in both empirical and computational results.

3.3 Experiment 2

Experiment 2 replicated the dot counting design of Experiment 1 using a more complicated quadrimodal distribution with the aim of providing a stronger contrast between the operation of the discrete and continuous priors. As such, the hypotheses of this experiment were identical to the first, expecting a greater range of responses in the more uncertain condition under a continuous system and a smaller number of responses under a discrete system, though the design was expected to be more diagnostic in separating these hypotheses in this case. In addition, this task also used

a larger sample size to provide more statistical power given the reasonably weak findings of the first experiment.

3.3.1 Method

Participants

Sixty University of Warwick students were recruited as participants in the experiment from the university's online SONA system in return for £6 in payment. The sample included 36 females and 24 males, while age ranged between 18 and 39 years, with a mean of 22.5. Payments were again unrelated to performance in the task.

Design

The design of Experiment 2 was identical to that of Experiment 1 with the exception of the underlying distribution: in place of the bimodal distribution, a quadrimodal distribution was used (illustrated in Figure 3.3).

Procedure

Experiment 2 used the same procedure as Experiment 1. Assignment to uncertainty conditions was again randomised and controlled to provide equal numbers in each group, meaning 30 participants were assigned to the 70% condition and 30 to the 95% condition.

3.3.2 Results

Data from Experiment 2 was analysed using the same procedure as Experiment 1, including the same exclusion criteria; while no participants were entirely removed from analysis in this task, an average of 2.33% ($\pm$ 0.71% 95% confidence intervals) of responses across participants fell more than 10 points outside of the displayed

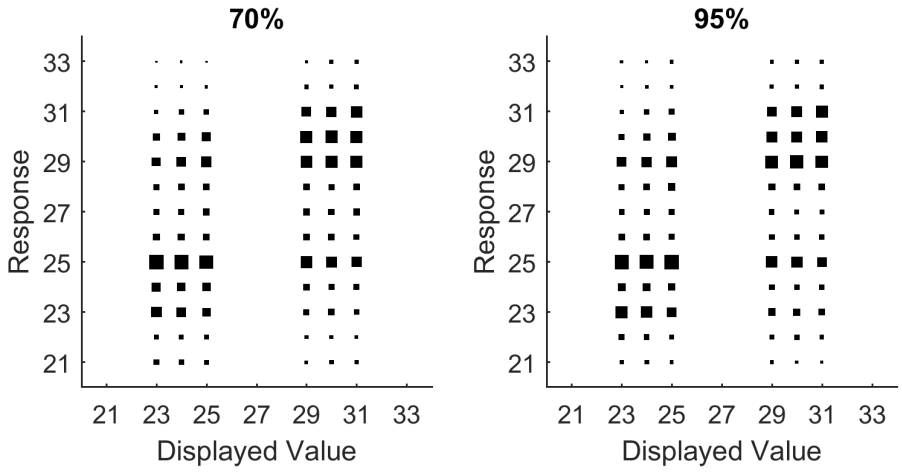

Figure 3.4: Conditional response distributions from the 70% and 95% uncertainty conditions of Experiment 2, where square size is proportional to the percentage of responses made to each displayed value.

range, and so were classified as response errors and eliminated from subsequent comparisons.

Figure 3.4 shows conditional response distributions from the two uncertainty conditions in Experiment 2. As with the previous experiment, both groups demonstrated reasonable acquisition of the true quadrimodal structure, again showing strong preferences for the modes of the distribution; participants' mean number of responses across the four modes was again significantly greater than the mean number of responses across the remaining five values, $t(59) = 7.01$, $p < .001$, $d = 1.61$. As in Experiment 1, however, participants in both conditions did also use unseen values in their responses, in this case including the unused values from the midrange of the distribution, again suggesting bleed-over in both groups.

Comparisons from the second experiment are summarised in Table 3.2. As in Experiment 1, the count of unique responses was found to be significantly higher in the 70% condition, $t(58) = 2.21$, $p = .031$, $d = 0.59$, showing a greater range in the more uncertain condition. Once again, however, no significant difference was found between the 70% and 95% groups in either the number of out-of-range responses, $t(58) = 0.53$, $p = .600$, $d = 0.14$, or the number of mid-range (zero-probability) responses, $t(58) = 0.80$, $p = .425$, $d = 0.21$, though these were again both higher in

the 70% group.

These results therefore correspond with the findings of the first experiment: participants in the high-uncertainty condition used a wider range of values in their responses, but did not demonstrate a reliable increase in the use of unshown values over those in the low-uncertainty condition. This again provides limited evidence for the use of a continuous mixture prior, with components seemingly becoming broader under uncertainty, thereby covering more potential values, but not necessarily relying on those values. However, while both experiments may offer weak demonstrations of continuous effects in isolation, by replicating the observed effects in two separate tasks, these results combine to provide more reliable evidence, suggesting behaviour in these tasks was in fact based on the use of a continuous numerical system.

The collected empirical data then provides a reasonable qualitative indication of the numeric format underlying estimation based on a theoretical contrast of the behaviour of the two considered priors: reactions to uncertainty better match the predictions of a continuous system than a discrete system. To supplement these findings, however, behavioural data was next directly compared with computational models of estimation for a quantitative assessment of the fit of both the Gaussian mixture and categorical priors to the collected data. This also allowed for an examination of general behavioural trends across all participants beyond the distinction between the two uncertainty conditions of these empirical contrasts, offering an alternate exploration of the processes underlying behaviour in these experiments.

Condition	70%	95%
Unique response count	17.0 ($\pm$ 1.57)	14.7 ($\pm$ 1.53)
Out-of-range responses	66.3 ($\pm$ 24.7)	56.2 ($\pm$ 30.2)
Mid-range responses	69.2 ($\pm$ 23.7)	55.6 ($\pm$ 25.4)

Table 3.2: Mean measures with 95% confidence intervals from the two uncertainty conditions of Experiment 2.

3.4 The Uncertain Estimation Model

In order to investigate the underlying processes used in the experimental tasks, we developed a perceptual estimation model which was able to use either a continuous or discrete prior format while other model elements remained identical. This drew on existing clustering models in which observations are assigned to subgroups based on similarities in features as well as subgroup size, most notably the Rational Model of Categorisation (RMC) by Anderson (1991) which uses Bayes' rule to approximate the ideal partition of items. As noted above, such systems have previously been successfully applied to numerosity (Gershman & Niv, 2013), as well as language comprehension (Goldwater et al., 2009) and causal reasoning (Buchsbaum et al., 2015).

The present model therefore considers potential assignments of observations to subgroups based on perceptual data, trial feedback and prior experience in the task, creating a set of clusters which can be aggregated to provide a representation of the true external distribution. The format of these clusters however is dependent on the utilised prior, here limited to the previously noted categorical and Gaussian mixture priors to contrast discrete and continuous numerical structures. The model is therefore nearly identical to the definitions of the RMC given by Anderson (1991) for discrete and continuous dimensions, here adapted to infer a physical feature for a set of cluster members rather than a category label. This model was named the 'Uncertain Estimation Model', or UEM.

On each estimation trial, the model determines the probability of each potential value in each potential cluster generating both the observed perceptual data and the given feedback value across all possible partitions of past observations:

$$p(S_t|X_{1:t},F_{1:t}) = \sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(S_t,S_{1:t-1},Z_t,Z_{1:t-1}|X_{1:t},F_{1:t}) \tag{3.1}$$

where t is the current trial, $S_{1:t-1}$ is a vector containing the dot counts $S_1,S_2,...,S_{t-1}$, $Z_{1:t}$ is a vector containing the cluster indices $Z_1,Z_2,...,Z_t$, $X_{1:t}$ is a vector containing

the perceptual data $X_1, X_2, ..., X_t$ and $F_{1:t}$ is a vector containing the feedback values $F_1, F_2, ..., F_t$. This can be broken down to isolate the probability of the proposed value generating the observed perceptual and feedback data:

$$p(S_t|X_{1:t}, F_{1:t}) \propto$$
$$\sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(X_t|S_t)p(F_t|S_t)p(S_t|S_{1:t-1}, Z_{1:t})p(Z_t|Z_{1:t-1})p(S_{1:t-1}, Z_{1:t-1}|X_{1:t-1}, F_{1:t-1}) \tag{3.2}$$

This equation is composed of five elements to be calculated: first, $p(X_t|S_t)$ notes the probability of the observed perceptual stimulus X_t given the potential value S_t, where X_t is an estimate of the perceptual stimulus sampled from a lognormal distribution with mean equal to the logarithm of the true dot count v_t and fixed variance σ_l^2 based on assessment of the observer's discrimination ability:

$$p(X_t|v_t) = logN(X_t; \ln(v_t), \sigma_l^2) \tag{3.3}$$

This estimate is then compared with each considered value using a second lognormal distribution with mean equal to the logarithm of the considered value and equal variance:

$$p(X_t|S_t) = logN(X_t; \ln(S_t), \sigma_l^2) \tag{3.4}$$

Secondly, $p(F_t|S_t)$ notes the probability of the feedback score given the proposed value, allowing for the consideration of uncertainty in feedback information; this treats feedback as a perceptual feature of the trial rather than a definitive label, assessing the fit of this information to the considered value. For the purposes of simplicity, this uses a single parameter to reflect the assumed reliability of trial feedback:

$$p(F_t|S_t) = \begin{cases} c_f & \text{where } F_t = S_t \\ \\ \frac{1-c_f}{n_v-1} & \text{otherwise} \end{cases} \tag{3.5}$$

where c_f is the feedback accuracy parameter, fixed across all trials, and n_v is the number of values considered for S_t.

Thirdly, $p(S_t|S_{1:t-1}, Z_{1:t})$ notes the probability of the proposed value given the partition suggested by $S_{1:t-1}$ and $Z_{1:t-1}$ and the proposed cluster Z_t. This term therefore introduces the distinction between continuous and discrete structures, as this affects the generated partition.

3.4.1 Discrete Format

For the discrete form, a count of matching observations is used:

$$p(S_t|S_{1:t-1}, Z_{1:t}) = \frac{n_s}{n_z} \tag{3.6}$$

where n_s is the count of observations in cluster Z_t with value S_t and n_z is the total membership of cluster Z_t; this distribution therefore becomes binary for non-empty clusters due to the uniformity of their membership, being 1 where S_t matches the value of these members and 0 elsewhere. For new potential clusters without any members, this instead uses a uniform prior across the considered values of S_t. This distribution therefore matches the definition used by the RMC for likelihood values using discrete dimensions where the prior expectancy parameter used by the RMC (α) approaches zero.

3.4.2 Continuous Format

For the continuous form, a Gaussian mixture is used, computing the mean and variance of the cluster distribution given its currently assigned members as well as an assumed prior mean and variance independent of any observations. This follows the definition given by Anderson (1991) for likelihoods using continuous dimensions, in which an inverse chi-squared distribution to provide an estimate of the variance:

$$\sigma^2 \sim \beta_0 {\sigma_0}^2 {\chi_{\beta_0}}^2 \tag{3.7}$$

where ${\sigma_0}^2$ is the prior variance and β_0 refers to the confidence in this prior variance, while the mean uses a Gaussian distribution:

$$\mu|\sigma \sim N\left(\mu_0, \frac{\sigma}{\sqrt{\lambda_0}}\right) \tag{3.8}$$

where μ_0 is the prior mean and λ_0 is the confidence in this prior mean (note that the second parameter of this distribution is the standard deviation rather than the variance). The use of these two distributions then results in a t-distribution describing the probability of value S_t in the given cluster (again, the second parameter of this t-distribution is the standard deviation rather than the variance):

$$p(S_t|S_{1:t-1}, Z_{1:t}) = t_c(S_t; \mu_i, \sigma_i\sqrt{1+1/\lambda_i}) \tag{3.9}$$

The parameters of this distribution are calculated according to the proposed membership of the target cluster in the currently assumed partition, combining the prior mean μ_0 and variance ${\sigma_0}^2$ with the observed mean $\bar{x}$ and variance s^2 using the confidence values β_0 and λ_0:

$$\beta_i = \beta_0 + n_z \tag{3.10}$$

$$\lambda_i = \lambda_0 + n_z \tag{3.11}$$

$$\mu_i = \frac{\lambda_0\mu_0 + n_z\bar{x}}{\lambda_0 + n_z} \tag{3.12}$$

$$\sigma_i^2 = \frac{\beta_0\sigma_0^2 + (n_z-1)s^2 + \frac{\lambda_0 n_z}{\lambda_0+n_z}(\mu_0-\bar{x})^2}{\beta_0 + n_z} \tag{3.13}$$

Fourthly, $p(Z_t|Z_{1:t-1})$ is a Chinese Restaurant prior (Aldous, 1985; Pitman, 2002) describing the probability of the observation being assigned to cluster Z_t

based on the size of that cluster, following the format of Anderson (1991):

$$p(Z_t|Z_{1:t-1}) = \begin{cases} \dfrac{cn_z}{(1-c)+cn} & \text{if } Z_t \text{ is old} \\ \\ \dfrac{(1-c)}{(1-c)+cn} & \text{if } Z_t \text{ is new} \end{cases} \tag{3.14}$$

where n_z is the number of observations in cluster Z_t in the current partition, n is the total number of assigned observations and c is a coupling parameter describing the probability of two items being grouped together independent of any other observations.

Finally, $p(S_{1:t-1}, Z_{1:t-1}|X_{1:t-1}, F_{1:t-1})$ describes the probability of the currently assumed partition given by $S_{1:t-1}$ and $Z_{1:t-1}$, which is equal to the product of the probability of each past observation's assignment to the partition as defined by Equation 3.2.

Once the probability of each potential permutation has been calculated, these values can be used to generate the predictive probability of any value appearing in the next trial by aggregating over the individual distributions of each potential partition:

$$p(S_{t+1}|X_{1:t}, F_{1:t}) = \sum_{S_{1:t}} \sum_{Z_{1:t+1}} p(S_{t+1}, S_{1:t}, Z_{t+1}, Z_{1:t}|X_{1:t}, F_{1:t}) \tag{3.15}$$

$$\propto \sum_{S_{1:t}} \sum_{Z_{1:t+1}} p(S_{t+1}|S_{1:t}, Z_{1:t+1}) p(Z_{t+1}|Z_{1:t}) p(S_{1:t}, Z_{1:t}|X_{1:t}, F_{1:t}) \tag{3.16}$$

Similarly, the UEM is able to calculate the probability of the responses made by participants in the present experimental procedure, where estimates are given based on the perceptual stimulus before receiving feedback, by simply omitting the feedback element from Equation 3.2:

$$p(S_t|X_{1:t}, F_{1:t-1}) = \sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(S_t, S_{1:t-1}, Z_t, Z_{1:t-1}|X_{1:t}, F_{1:t-1}) \tag{3.17}$$

$$\propto \sum_{S_{1:t-1}} \sum_{Z_{1:t}} p(X_t|S_t)p(S_t|S_{1:t-1},Z_{1:t})p(Z_t|Z_{1:t-1})p(S_{1:t-1},Z_{1:t-1}|X_{1:t-1},F_{1:t-1}) \tag{3.18}$$

3.4.3 Details of Model Approximations

While the above equations do provide a calculable formula, by considering all possible permutations of past cluster and value assignments, the full version of the model would quickly become intractable at even a moderate number of observations. As such, this full solution is approximated by reducing the number of considered permutations to a set of samples using particle filtering. This process makes use of a fixed number of 'particles', each containing a possible permutation of cluster and value assignments for past trials at that point in time (Griffiths, Sanborn, Canini, Navarro, & Tenenbaum, 2011). Following a new observation, the model considers only the assignments of that observation which are consistent with current particles, calculating the probability of the assignment according to:

$$p(S_t,Z_t|X_{1:t},F_{1:t}) \approx \frac{1}{n_i}\sum_i p(X_t|S_t^{(i)})p(F_t|S_t^{(i)})p(S_t^{(i)}|S_{1:t-1}^{(i)},Z_{1:t}^{(i)})p(Z_t^{(i)}|Z_{1:t-1}^{(i)}) \tag{3.19}$$

where $S_{1:t-1}^{(i)}$ and $Z_{1:t-1}^{(i)}$ represent the value and cluster assignments in particle i, and n_i is the number of particles. The equations for each of these components are therefore identical to those given above restricted to the partition held by the particle under consideration. Once the probability of each valid assignment has been calculated, these probabilities are then used to stochastically sample new partitions including the latest observation to be held as the new particles for the next trial.

Similar processes can then be performed for both prediction and response selection, again restricting the considered permutations to those currently held in the particles; as such, the predictive distribution becomes:

$$p(S_{t+1}|X_{1:t},F_{1:t}) \approx \frac{1}{n_i}\sum_i p(S_{t+1}^{(i)}|S_{1:t}^{(i)},Z_{1:t+1}^{(i)})p(Z_{t+1}^{(i)}|Z_{1:t}^{(i)}) \tag{3.20}$$

replacing Equation 3.16, while the response distribution becomes:

$$p(S_t|X_{1:t},F_{1:t-1}) \approx \frac{1}{n_i}\sum_i p(X_t|S_t^{(i)})p(S_t^{(i)}|S_{1:t-1}^{(i)},Z_{1:t}^{(i)})p(Z_t^{(i)}|Z_{1:t-1}^{(i)}) \tag{3.21}$$

replacing Equation 3.18.

In addition to the particle filter, the model included a second approximation within the perceptual distribution of Equation 3.4: to make computation more tractable, the sampled value X_t was replaced with the true value v_t, so assuming perceptual samples were perfectly accurate:

$$p(X_t|S_t) = logN(X_t;\ln(v_t),\sigma_l^2) \tag{3.22}$$

replacing Equation 3.4. While this does remove some noise from the estimation system, this can be subsequently reinserted by sampling responses from the distribution given by Equation 3.21 rather than simply taking the maximum, an approximation which has previously been found to be successful (e.g. Sanborn, Mansinghka, & Griffiths, 2013).

Finally, for the purposes of fitting the UEM to actual behaviour, the response distribution was further edited to include two additional elements: first, the distribution is raised to an exponent to allow the model to interpolate between probability matching and maximisation, and second, the response distribution is combined with a uniform background distribution to emulate potential noise in response selection:

$$p(R_t|X_{1:t},F_{1:t-1}) = (1-w_b)\frac{p(S_t|X_{1:t},F_{1:t-1})^e}{\sum p(S_t|X_{1:t},F_{1:t-1})^e} + w_b U(v_1,v_2) \tag{3.23}$$

where R_t is the potential response, e is the response exponent, w_b is the weight applied to the background distribution and v_1 and v_2 provide the range of values considered in the uniform distribution. Responses can then be drawn from this distribution using various methods, with the resulting feedback being used to update the representation using the above method. For the purposes of this study, however, no fixed sampling method is defined, with this distribution instead being used to

provide the probability of a given participant response.

3.4.4 Model Comparison

The discrete and continuous forms of the UEM were compared with the experimental data from both Experiments 1 and 2 using a grid point search across the four parameters shared by the two models to determine the best fit to the collected data. This was used in place of more traditional gradient descent functions due to potential issues with such methods for clustering models: the likelihood function of these models is often highly complex, leading gradient descent functions to become fixed at local maxima rather than the global maximum. The search ran across the four parameters shared by the two models: the coupling parameter c, response exponent e, feedback confidence c_f and background weight w_b. Considered values were: for c, 0.1 to 0.9 in steps of 0.1; for e, 0.1, 0.25, 0.5, 1, 1.5 and 2; for c_f, 0.1, 0.3, 0.5, 0.7 and 0.95 (capturing the stated accuracy in the 95% condition); and for w_b, 0.01, 0.1, 0.3, 0.5, 0.7 and 0.9. The models also used a particle filter as described above to aid computation, both set to use five particles.

In order to make the modelling more computationally tractable, the prior parameters unique to the cUEM (μ_0, ${\sigma_0}^2$, β_0 and λ_0) were fixed across model fits. The values of these parameters were set according to the range of displayed dot counts following the format of Anderson (1991) in which the prior mean is set at the midpoint of the range (Experiment 1: 27.5; Experiment 2: 27), and prior variance is set at a quarter of the range squared (Experiment 1: 5.06; Experiment 2: 4), while confidence values for both these parameters are set at one. However, in order to allow for the previously described emulation of categorical components by the Gaussian mixture prior, the prior variance and confidence values were edited to provide a narrower initial form; prior variance was therefore set at a twentieth of the range squared (Experiment 1: 0.20; Experiment 2: 0.16), while the associated confidence value λ_0 was set at 0.01, determined through limited likelihood testing using manual adjustments of these parameters on a subset of the data. While

these manipulations were limited, these were considered as full parameters for the purposes of calculating complexity penalties in subsequent measures. As such, the dUEM was defined as having four free parameters, and the cUEM was defined as having six.

Both models were then fit to each participant individually by providing the respective model with the observed dot counts in matching order for partitioning, calculating the response distribution (given in Equation 3.23) and taking the resulting probability of the participant's response for that trial. These trial probabilities were then converted into log values and summed to produce a log likelihood figure for each participant at each grid point of each model. For greater reliability in fit, each grid point was repeated three times to provide an average log likelihood; this was limited at a reasonably low count to aid computation, though additional comparisons from a subset of parameters between likelihoods averaged across either 3 or 50 trials found high correlations (dUEM: r = 0.962; cUEM: r = 0.950), suggesting this was still a reasonably accurate estimate.

Maximum log likelihood values for each participant from each model were then converted to Akaike information criterion (AIC, Akaike, 1974) and Bayesian information criterion (BIC, Schwarz, 1978) values for further comparison due to the differing number of parameters between the models. These measures both provide an adjusted measure of model fit controlling for model complexity, with lower values indicating a better fit. Both measures were also used to calculate weights for the given comparison between the cUEM and dUEM, providing an estimate of the posterior probability of each model assuming equal priors (Wagenmakers & Farrell, 2004). For ease of presentation, the following analysis focuses primarily on BIC scores as the more conservative measure, with AIC results being noted where these scores suggest a qualitative difference in outcome, while full AIC results are listed in the appendix.

The grid point structure allows for both global and individual fitting, either aggregating likelihoods across participants at each grid point assuming a common set of parameters within each experiment, or calculating a maximum likelihood

value for each participant assuming differences in parameters between individuals, converting that maximum to a BIC score and then aggregating the resulting values. Both AIC and BIC measures from both experiments do however show fits are substantially better when using individual parameters; as such, the remainder of the comparison uses aggregates of individual best fits by summing BIC scores from each participant, with global fits being listed in the appendix.

Results of the modelling comparison with Experiment 1 data are summarised in Table 3.3. Across all participants, the cUEM had a better fit to the data by summed BIC scores than the dUEM, though a nearly equal division was observed between the number of participants best fit by the dUEM (20) and the cUEM (19). When separated by uncertainty condition, the cUEM provided a better fit to the 70% group, accounting for 11 of the 19 participants, while the dUEM had a better fit to the 95% group, accounting for 12 of the 20 participants. This does not necessarily mean however that different priors were used between conditions, as participants are unlikely to select a prior according to the experimental manipulation. Instead, this may reflect the above suggestion that behaviour appears more discrete where feedback is more reliable as this is where continuous components are best able to emulate discrete structures, with the better fit of the discrete model then being a result of this emulation; while this suggestion is supported by summed BIC scores from the two groups, a chi-squared test found no significant difference in the ratio

Comparison	Model	Parameters	MLL	BIC	*w*(BIC)
Individual	dUEM	4	-43720	88409	0
	cUEM	6	-43461	88377	1
70%	dUEM	4	-22335	45142	0
	cUEM	6	-22112	44933	1
95%	dUEM	4	-21385	43267	1
	cUEM	6	-21349	43444	0

Table 3.3: Modelling results from Experiment 1, where MLL is the summed maximum log likelihood for that model across participants, BIC values are summed from individual best fits, and *w*(BIC) is the weight of the BIC score for the given comparison between the discrete and continuous models, approximating the posterior probability of each model assuming equal priors.

of participants best fit by the two models between uncertainty conditions, $\chi^2(1) = 1.25$, $p = .264$.

AIC scores offer almost identical qualitative results, with the only notable difference being in the number of participants best fit by each model, which shows a significant difference in ratio between the two certainty conditions, $\chi^2(1) = 4.50$, $p = .034$, suggesting behaviour did appear more continuous in the high-uncertainty condition. Margins between AIC scores also demonstrate a stronger support for the cUEM, showing greater advantages where this model fits better and a narrower difference in the 95% condition where the dUEM remains ahead. Such results are attributable to the reduced cost of complexity in AIC scores, more closely reflecting the difference in raw likelihood despite the different parameter counts of the two models. This is notable given that the current comparisons did not take full advantage of the greater complexity of the cUEM, as the additional parameters of this model were in fact fixed across the comparison, but were treated as variable given the initial manual manipulations of variance and confidence to allow for narrower components. This does however mean that the cUEM performed better even under the harsher complexity costs of the BIC measures, providing further support for this prior.

In contrast with the first experiment, the cUEM displayed a greater advantage in the number of participants best fit by each of the models in Experiment 2, accounting for 37 of the 60 participants; this is further displayed in the summed BIC scores, which show the cUEM had a better overall fit to the data, detailed in Table 3.4. Separated by group, summed BIC scores again found the cUEM to have a better fit in the 70% condition, accounting for 19 of the 30 participants, though this model now also better fits the 95% condition, accounting for 18 of the 30 participants. As with the first experiment, this difference in ratio between the two groups was found to be non-significant, $\chi^2(1) = 0.07$, $p = .791$.

AIC scores meanwhile again show almost identical results, though with slight differences in the ratios of participants best fit by each model, again seemingly showing greater support for the cUEM where the penalty for complexity is

Comparison	Model	Parameters	MLL	BIC	*w*(BIC)
Individual	dUEM	4	-66731	134953	0
	cUEM	6	-65994	134225	1
70%	dUEM	4	-34740	70225	0
	cUEM	6	-34338	69796	1
95%	dUEM	4	-31991	64728	0
	cUEM	6	-31655	64429	1

Table 3.4: Modelling results from Experiment 2.

less severe.

Results from both model comparisons therefore suggest that a Gaussian mixture prior was more likely to be used in their respective tasks than a categorical prior, so supporting the apparent continuous effects observed in the empirical contrasts. This in fact extends to the one notable difference between the first and second model comparisons: behaviour in Experiment 2 is better fit by the cUEM even in the 95% condition where lower uncertainty was suggested to allow behaviour to appear more discrete. Such a difference is attributable to the greater complexity of the quadrimodal distribution used in the second task, being more difficult to emulate using continuous structures; this makes continuous effects more apparent even where feedback is more reliable. Even so, it is notable that a substantial number of participants in both tasks were better fit individually by the dUEM. As such, there is some remaining ambiguity as to the prevalence of the continuous prior in such estimates, with potential individual differences in prior format across participants. Such a suggestion will however require further qualitative contrasts in future work to more definitively identify if these are in fact reliable differences in numerical format between individuals, particularly where general model fitting measures are highly supportive of the continuous system overall.

These comparisons therefore do indeed further support the above empirical findings: while behavioural data in both experiments demonstrates qualitative evidence of a continuous representation of past numeric experience, this is now reinforced quantitatively by model fitting, providing greater confidence in this conclusion. This highlights the difference between the empirical and computational

comparisons used here: while empirical contrasts focus on the differences in behaviour between the two uncertainty conditions, which may be limited in scope, the model comparison is able to examine wider behavioural patterns across all participants, identifying a trend towards continuous behaviour common to both groups.

3.5 Discussion

The above sections provide evidence from two experiments of a continuous numerical system underlying discrete estimates which reacts to uncertainty by simplifying the held representation using rational categorisation principles: in both tasks, responses become more varied when feedback was less reliable, indicating a broadening of Gaussian components. This is further supported by comparisons with computational models of estimation: in both experiments, behaviour was better fit by a Gaussian mixture prior over a categorical prior, providing a second source of evidence for the use of a continuous prior format, though further examinations may be required to determine the prevalence of this system across individuals. Both discrete and continuous estimates therefore appear to share a common continuous estimation system able to adjust the formed representation to best meet the needs of the task.

These results also provide greater insight into the findings of Sanborn and Beierholm (2016), further clarifying the process by which numerical estimates are made: through direct comparisons with full models of the estimation process, these results suggest that learners are able to acquire complex multimodal distributions through the use of a highly flexible continuous numerical system able to emulate such detailed structures. This then allows for the appearance of the use of discrete numerical formats in such tasks despite actually being based in continuous systems, offering a new window into the results of Sanborn and Beierholm (2016): the apparent use of discrete priors in that study may in fact be the result of a continuous system emulating the narrower component format of a truly discrete distribution. This may be attributable to aspects of the design of that study which facilitated

such emulation: for example, the range of values displayed in the task was reasonably small in comparison to other studies (e.g. Gershman & Niv, 2013), potentially encouraging the use of a set of narrow components to provide better discrimination. Alternatively, the use of definitive feedback may have avoided potential noise in value assignment which could broaden components: the results of Experiment 1 find behaviour better matches the use of discrete formats when feedback is more reliable, though this was not replicated in Experiment 2, while the number of participants best fit by the dUEM did not significantly differ between uncertainty conditions in either task. Even so, this effect could be more pronounced for the exact feedback used by Sanborn and Beierholm (2016) as participants are given no reason to believe this information is inaccurate, further narrowing components for a closer emulation of discrete structures.

The present findings therefore depict a highly flexible estimation system in which any formed representation and resulting behaviour are highly sensitive to the scenarios that produce them. This allows the system to acquire more complex distributions such as those used in the present experiments: without such a representation, learners would not be able to accurately capture such forms. This can in fact be demonstrated by lesioning the present models to remove their respective priors; if the model does not store any experiences in memory, then the learner is unable to update their beliefs, and decisions are based solely on perceptual evidence (detailed in the appendix). Such lesioning generates a drop in estimated accuracy for both discrete (51.1% vs. 42.7%) and continuous (52.9% vs. 45.5%) formats, illustrating the benefits to learning provided by such a system. In addition, as seen in the above experiments, this flexibility also allows the learner to account for uncertainty in the formed representation, further altering mental structures according to noise in the environment such as the unreliable feedback of the present designs. As such, these results help to demonstrate the power of a rational system in this task, utilising both direct observations and background knowledge to build a mental representation which accurately captures both external patterns and their surrounding context.

Such results also offer a notable correspondence with the wider literature on numerosity in which numbers often appear to be considered within a continuous format: even when presented symbolically, behaviour seems to suggest numerical values are treated continuously, showing greater confusion between similar values (Moyer & Landauer, 1967; Spelke & Tsivkin, 2001; Dehaene & Marques, 2002). The present study may then further contribute to the suggestion that learners rely primarily on approximate number systems when dealing with numerical values, translating the output of such systems into discrete figures when required (Izard & Dehaene, 2008). This links to the concept of 'number sense' (Dehaene, 2011), an innate understanding of numerosity displayed independently of the standard symbolic numerical system, as evidenced by its use by not just adult learners, but also infants (McCrink & Wynn, 2004) and animals (Flombaum, Junge, & Hauser, 2005; Ditz & Nieder, 2016). The apparent use of continuous structures across numerical tasks may then reflect a general reliance on this number sense, utilising a more fundamental numerical system where possible and converting this to symbolic formats as needed rather than directly working in a purely symbolic format learned in later life. What is more, the current results demonstrate that despite being a more primitive system, these structures can still enable efficient learning under the right circumstances: within the framework of a rational clustering process, continuous structures can be used to represent reasonably complex distributions, particularly where their inherent flexibility can be exploited.

In addition to the format of numerical information, the present distinction between discrete and continuous structures also demonstrates the impact of this structure on behaviour through the application of simplicity: the two priors provide almost directly opposing reactions to uncertainty, with one reducing the number of considered responses in order to simplify response selection, and one reducing the number of response regions but allowing for more potential values. The apparent use of continuous numerical structures therefore carries distinct behavioural implications, suggesting a greater reliance on prior expectations where feedback is less reliable. This then draws estimates towards previously expected values, but

without necessarily disregarding such information; returning again to the example of counting people in a room, if the observer receives a potential count from another individual that is viewed as unreliable, they are unlikely to store that figure in memory, but may use a similar number that falls between the feedback figure and their own prior expectations. This is in contrast to the more extreme process of the discrete model, where unreliable feedback may be completely abandoned in favour of prior values. Such a distinction is important given that real-world estimates are rarely followed by definitive feedback; even where such information is provided, this can be vague, or from an untrustworthy source. This also illustrates the broader importance of understanding the form of our representations, as slight differences in structure can have substantial effects on behaviour. As such, any interventions into such systems must consider what structures people may hold in order to provide meaningful results; in the current case, this applies primarily to methods that may encourage more accurate learning of real-world distributions, though this concept applies to any action based on internal mental representations.

It should be noted again however that the present Bayesian models were used as descriptions of behaviour to facilitate the comparison between discrete and continuous prior formats, and do not necessarily reflect the processes used by actual learners when making numeric estimates. This also places the current models at the computational level of analysis (Marr, 1982), offering high-level principles for behaviour rather than any specific algorithmic mechanism that may be used by actual learners. Even so, BDT does remain a strong candidate for the true process: as previously noted, BDT provides a better account for the use of prior information than theories such as calibration (Sanborn & Beierholm, 2016), allowing for the acquisition of more complex distributions such as those used in the present study. In addition, existing work has offered a number of algorithms which could support Bayesian models such as these, most notably sampling methods (Gelman et al., 2013), which have been found to accurately account for human biases in a number of tasks (Sanborn, Griffiths, & Navarro, 2010; Griffiths, Vul, & Sanborn, 2012; Sanborn & Chater, 2016). The current results are not however able to definitively

determine the validity of the considered Bayesian models, meaning these models remain descriptive until more direct tests are performed.

Another caveat to these conclusions is that the current study is limited in the priors considered in the above model comparisons, focusing on only two particular systems to suit the contrast between continuous and discrete structures. While these systems do serve the present examination of numerical format in distribution learning, there are multiple other priors which could be investigated as alternatives to these processes, including more complex continuous or discrete systems which do not follow the same behavioural predictions used here; for example, discrete components which correlate with neighbouring values could emulate the generalisation pattern suggested for continuous structures whilst still using a discrete underlying numerical format. The present study does however remain primarily focused on the distinction between continuous and discrete numerical systems offered by existing numerical research, including the difference in generalisation between these systems stated in the introduction to this section where bleed-over is predicted by continuous but not discrete structures (Moyer & Landauer, 1967; Dehaene, 2011). The currently considered priors are therefore most appropriate to the aims of this study, while such alternative priors can be considered in future work.

Finally, one additional factor to consider in this study is the method by which uncertainty was manipulated in this design: in order to create doubt in the task feedback, true values were presented as answers given by a past participant, using that participant's reported accuracy rate as a measure of reliability. This therefore introduces a social information element to the task, as participants are made to consider the method by which these feedback values are generated. This is particularly notable given that previous research has found that learners may draw different inferences from observed data according to its origin: beliefs may differ when examples are chosen by a teacher to illustrate an idea (Shafto, Goodman, & Griffiths, 2014), or when samples are noted to exclude certain results (Hayes, Banner, & Navarro, 2017) compared to observation alone. While the current task is unlikely to have encouraged these particular higher level inferences, the origin of feedback remains a

consideration when determining how participants interpret this information during decision making: there are multiple potential methods of using feedback data with varying levels of complexity, ranging from a reasonably simplistic correct/incorrect dichotomy to a full model of the past participant's decision process. For the purposes of simplifying model fitting, the most basic of these forms were used in both of the present models, using a single parameter to reflect the probability of the feedback being accurate; future work on this subject may therefore wish to consider these alternate definitions in order to provide a more complete model of behaviour. Alternatively, similar tasks could make use of non-social manipulations of uncertainty to assess the impact of this factor on decision making.

3.5.1 Conclusion

The present study provides both empirical and computational evidence that discrete numeric estimates are built on continuous mental structures, displayed here via reactions to uncertainty: learners react to unreliable feedback by broadening their response regions, utilising the inherent flexibility of their representation to account for noise in the environment. This demonstrates not just the systems used within numerical estimation, but also the impact of these systems on both the distributions learned through this process as well as behaviour built on this representation. We therefore hope that this study can provide a basis for further examination of the mechanisms underlying numerical estimation, using additional experimental contrasts and more advanced computational models to offer greater insight into these systems, and so the wider representation of numerical information.

Chapter 4

Trial Replay in Learning Consolidation

In order to best direct behaviour, our representations should accurately reflect external structures, capturing the connections between items as well as their associated costs and benefits. Such a representation is not necessarily, however, built solely on direct observation, but could also involve the consolidation and potential revaluation of these structures outside of learning; new information could re-frame existing knowledge (or vice versa), or provide greater insight into concepts that were not previously fully understood. In this chapter, we examine how acquired representations are consolidated outside of direct learning using the Replay model, an extension to established associative learning models which permits for the revaluation of existing structures by revisiting past experiences. This involves three applications of this model to learning behaviour in various tasks, each investigating how such consolidation processes could assist in identifying associative structures which had previously been unnoticed. These experiments then examine the possible role of rehearsal in generating more complete mental representations, and so leading the learner towards more beneficial responses for the given situation.

4.1 Rehearsal in Associative Learning

Existing associative learning models have often focussed on the use of direct observations in forging associative connections between stimuli, a key example being the Rescorla-Wagner (RW) model, where the strength of an association is adjusted according to errors between predictions and observations (Rescorla & Wagner, 1972). While this describes the basic method by which an association can be acquired and extinguished, by basing learning only on direct observations of relevant stimuli, the RW model is unable to account for several learning phenomena which have been observed in actual behaviour. For example, the RW model suggests that an extinguished response should remain extinguished without further reinforcement, while in reality, such a response can in fact re-emerge at a later point given a sufficient interval, an effect known as spontaneous recovery (Rescorla, 2004). Similarly, prior exposure to a stimulus is not predicted to affect subsequent acquisition of a related association in the RW model as no learning is thought to occur in this period, whereas this can in fact slow later training using this stimulus, known as latent inhibition (Lubow, 1973). These effects, as well as others such as backwards blocking and backwards conditioned inhibition, are suggestive of a more complex learning system than is offered by the RW model, in which the mental representation that is ultimately formed is based upon the wider framework of all learning experiences rather than focusing on the most recent set of trials. It is therefore necessary to consider more complex learning models which are able to review and adjust learned structures to identify broader associative patterns in order to account for these more advanced learning behaviours.

One potential solution for these issues is to draw on existing concepts of rehearsal and consolidation used in theories of memory; such processes have been suggested to be key to the transition of recent experiences to long-term storage (Atkinson & Shiffrin, 1968; McGaugh, 2000; Ratcliff, 1990), particularly during periods of sleep (Stickgold, 2005; Born, Rasch, & Gais, 2006; Maquet, 2001). The application of such concepts to associative learning could then allow for learning

models which are able to revisit past training and use this rehearsal to further adjust the representation to reflect all experiences. Such a mechanism has in fact been offered in existing associative learning studies, with the suggestion that rehearsal of past learning experiences could allow for the alteration of acquired associative structures, potentially explaining some of the phenomena described above (Chapman, 1991; Ratcliff, 1990); for example, Gershman, Markman, and Otto (2014) inserted a short break between training and test in a retrospective revaluation design where new information suggests a previously trained response is no longer optimal, and found significantly greater levels of revision of previous preferences, indicating a role of offline rehearsal in adjusting perceived stimulus values. This is further supported by physiological evidence of such rehearsal from neurological studies: hippocampal cells associated with specific locations have been found to fire in similar sequences during training and rest, suggesting the neural rehearsal of learning experiences (Wilson & McNaughton, 1994; Euston, Tatsuno, & McNaughton, 2007; Davidson, Kloosterman, & Wilson, 2009). What is more, such neural replays appear to directly support the replanning of behaviour in similar revaluation designs, with increased activity being correlated with greater reversal of preferences (Momennejad, Otto, Daw, & Norman, 2018). Taken together, these findings then provide a substantial basis for a suggested use of consolidation processes within learning systems to build more complete mental representations of external structures.

The current study therefore aimed to investigate the role of rehearsal in updating associative representations, and the impact of such revision on related behaviour. This was implemented using the Replay model of associative learning (Ludvig, Mirian, Kehoe, & Sutton, 2017), a proposed extension to the Rescorla-Wagner model in which the learner is able to consolidate their experiences by mentally replaying past training trials outside of direct learning. This model draws on similar rehearsal processes to those suggested to assist with memory consolidation, here applied to the consolidation of associative learning structures: past trials can be replayed to review their events and outcomes, solidifying any observed associa-

tions without further training. This uses the same basic process as the standard RW model, adjusting the association between stimuli according to prediction errors; as such, the replayed trial is treated almost as a new experience, though it is given a lower weight than a true novel trial. Replays can, however, be run continually throughout rest, meaning longer breaks from training allow for greater levels of consolidation.

Importantly, any previous trial may be selected for replay, allowing for the rehearsal of both older and more recent trials. The Replay model is therefore able to adjust the associative structure to more accurately reflect all learning experiences, in contrast to the representation formed by the standard RW model, which focuses on the most recent events. This allows the Replay model to identify any broader patterns or structures that may be present in training, explaining several of the effects noted above: in the case of spontaneous recovery, the learner is able to replay both acquisition and extinction trials, so summating the two sets of trials and thereby producing an intermediate associative strength that is displayed in the recovery period. Similarly, latent inhibition is explained by a summation of the rewarded training trials and the unrewarded exposure trials during replay, suggesting an uneven pattern of reinforcement for the stimulus, thereby slowing the acquisition of the response. As such, through a minor adjustment to an established model allowing for the consolidation of learning experience, the Replay model appears to provide a more robust account for multiple aspects of associative learning.

This raises an interesting question: if the Replay model is able to identify broader patterns across all learning, could this allow for the discovery of associative structures that had not previously been fully understood? By replaying past events during periods of rest, learners could find rules or patterns that had not been noticed or acquired during training, apparently realising these structures without any new experience. This relates to the concept of insight, in which the solution to a problem is suddenly realised without any apparent conscious deliberation, often referred to as the 'Aha!' experience (Bowden, Jung-Beeman, Fleck, & Kounios, 2005). The Replay model may therefore provide a mechanistic explanation for insight-

like effects in standard associative learning paradigms, with past experiences being replayed unconsciously during breaks in training until complete solutions are eventually found.

The present study therefore aimed to investigate whether encouraging learners to perform replays of past training trials could allow for the discovery of previously unrealised learning structures; in the following sections, we present three experiments each examining this suggestion in differing tasks, beginning with an existing associative learning task, categorisation.

4.2 Experiment 1: Difficult Categorisations

Categorisations provide an existing associative learning design in which learners acquire associations between stimulus features and category labels through training on predetermined stimulus classifications, essentially learning a sorting rule for the given categorisation. While this could be taught through either description or experience, here we focus on the latter, with learners making stimulus categorisations and using feedback on the accuracy of their decision as a method of training, providing a store of observed trials which could be used in consolidation. The advantage of this task in the present study is that rules could be partially but not fully understood if sufficiently complex; learners could initially acquire a simplified version of true categorisation rules if this is accurate in most, if not all, cases. This partially acquired rule could then be refined through rehearsal of past trials, identifying remaining exceptions and revising the representation accordingly, so seemingly discovering the true rule without further training.

This first experiment therefore made use of a difficult categorisation task, in which learners are trained on complex, multidimensional categorisation rules until reaching a partial but incomplete understanding of the underlying structure; this drew on the hierarchy of basic categorisations provided by Shepard et al. (1961), in which items with three binary dimensions are organised into two equal groups, with the so-called 'Type IV' categorisation using all three stimuli dimensions (illustrated

in Figure 4.1) taking longest to be fully acquired. Interrupting this training should therefore provide a simple method for generating partial acquisition, as the learner should have some sense of the underlying rule, but is unlikely to have reached perfect performance. This can be measured according to the learner's accuracy over a subset of recent trials: if accuracy is above a certain criterion but not yet at ceiling, the learner can be said to have partially acquired the rule, and training can be ceased. The learner can then be given an opportunity to replay past trials, consolidating the partially acquired rule and so potentially leading to improved categorisation performance in a later test period. This follows the design of Gershman et al. (2014), using a short break between training and test to allow for a period of rehearsal, which was observed to generate greater levels of revaluation.

This therefore provided the basic design of the present experiment: categorisations were trained until meeting a performance criterion, before a break in training to allow for replays of the training trials, followed by further categorisation trials to assess any change in performance. However, in order to verify a replay benefit in such a task, it must be ensured firstly that learners are in fact performing replays during the break, and secondly that the performance of replays leads to a greater performance benefit compared to time away from the task alone. As such, a manipulation was required to encourage the replay of one set of categorisation trials over a comparable control. Such a manipulation is dependent, however, on the method by which experiences are selected for replay; if replays are used to help build more accurate representations of observed associative structures, there is likely to be some mechanism which prioritises certain experiences for rehearsal. This prioritisation would reduce the potentially vast collection of learning memories to those which are most relevant to the current situation. For example, replay could focus on the most recent trials under the assumption that temporally proximal events are likely to be more similar, or on trials with the most surprising outcome, examining the reasons for past expectation errors.

While these are valid possibilities, one particularly strong candidate is the context of learning; context is not only a clear, salient cue that could easily be

attached to trial memory, but has also been previously used in a similar manner to promote the consolidation of memory in sleep, both in humans (Rudoy, Voss, Westerberg, & Paller, 2009) and animals (Bendor & Wilson, 2012). Placing the learner back in a training context during a break in learning could then prioritise the replay of trials from that context over other categorisation trials, leading to greater consolidation of the rule associated with that context.

The following experiment therefore examined a potential replay benefit in a difficult categorisation task using a manipulation of training context: participants completed two sets of Type IV categorisation trials in distinct contexts until performance was above chance. This was followed by a break in training where the learner was placed in one of the two training contexts to encourage a replay bias, followed by further categorisation trials to assess subsequent performance. Based on the potential rehearsal benefit suggested by the Replay model, two hypotheses were examined in this task: firstly, that categorisation performance for the cued rule would be higher following the break, and secondly, that performance would be higher for the cued rule compared to the uncued rule.

4.2.1 Method

Participants

Eighty-four participants were selected from a University of Warwick undergraduate psychology class as part of a course requirement. The sample included 77 females and 7 males, while age ranged between 18 and 32 years, with a mean of 19.1.

Design and Materials

Two Type IV categorisation sets were used in the experiment, following the three binary dimension structure of Shepard et al. (1961); each set was therefore made up of eight stimuli divided into two pre-set categories of four. Both sets made use of the same dimensions for the stimuli (shape, colour and size), but used different values for these dimensions (squares/triangles and circles/hexagons, black/white

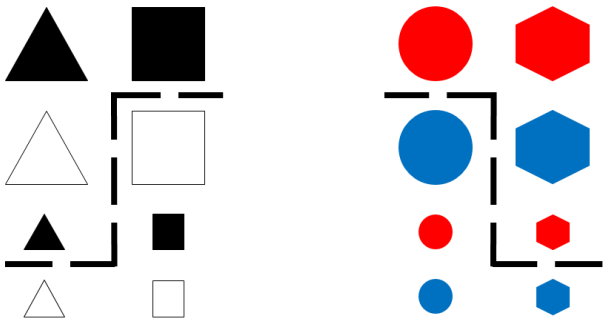

Figure 4.1: The two stimulus sets used in Experiment 1. Dotted lines demonstrate two potential Type IV rule boundaries.

and blue/red, large/small), illustrated in Figure 4.1.

In order to encourage the learning of a general categorisation rule rather than memorisation of individual stimuli, colour and size received slight variations to create the illusion of a wider stimuli set, converting the discrete labels given above into ranges and randomly sampling a value from these ranges under uniform probability for each presented stimulus. For size, this range represented the width in pixels of the shape (50-100 for small, 200-250 for large). For colour, this represented the deviation from the base Red-Green-Blue colour code of the stimulus, ranging between 0 and 75 points; for black and white shapes, this value was consistent to all colour dimensions (for example, a black stimulus with base colour code [0 0 0] may be presented with the altered colour code [50 50 50]), while for red and blue shapes, the key colour dimension remained constant (for example, a red stimulus with base colour code [255 0 0] may be presented with the altered colour code [255 60 60]).

Four potential permutations of the Type IV structure were available for each set; this was randomly selected at the start of each run of the experiment, though the two sets were prevented from sharing the same permutation.

Trial context was composed of two elements: a background image of a particular location to act as the training environment, and a coloured rim around the screen. These contexts were generated at the start of the experiment by randomly selecting an image and rim for each of the two categorisation rules. An example slide from the categorisation task is shown in Figure 4.2.

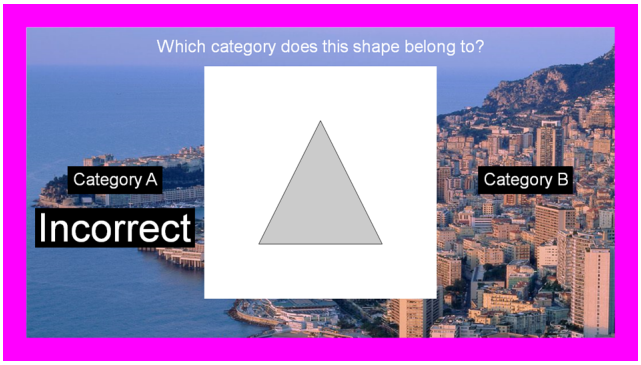

Figure 4.2: An example slide from Experiment 1, showing an incorrect categorisation response for the target stimulus.

Procedure

Participants began the experiment with training on the two categorisation rules; in each trial, a stimulus from one set was shown in the centre of the screen, along with two cues for the two possible classifications, with the contextual elements acting as a background. The participant was asked to identify which of two categories the shape came from by pressing the corresponding key on the keyboard. After making their response, feedback was provided by showing either 'Correct' or 'Incorrect' on screen, before advancing to the next trial. Training was divided into interlaced blocks of trials from each rule set to reduce the impact of order effects; participants therefore classified all eight stimuli from a set before switching to the alternative rule, changing the context, stimuli set and category labels.

After completing two blocks of a rule, the participant's performance on that categorisation was assessed. To ensure the participant had some understanding of the rule, accuracy had to be significantly greater than chance across the last sixteen trials to advance. A performance criterion was therefore set using a binomial distribution to provide the minimum number of accurate categorisations across this period that would qualify as significantly above chance ($p < 0.05$) if choices were random; this provided a criterion of 12 correct responses in the last 16 trials, for an accuracy rate of 0.75. Upon meeting this criterion, training for that rule was terminated; however, to prevent forgetting of the rule, blocks of the completed rule

moved from eight trials to one trial to continue to provide a reminder of the acquired rule.

Once criterion was met on both categorisation rules, the experiment advanced to the break phase to allow for replays. During the break, one context was randomly selected and shown on screen to encourage replays of the trials occurring in that context. In order to maintain attention towards the context during the break, participants were asked to complete a simple vigilance task, clicking Xs as they appeared on screen at 3 to 6 second intervals. The break lasted for three minutes before moving to the next phase, matching with the break length of Gershman et al. (2014).

Following the break, participants moved to the test phase, performing a further two eight-trial blocks of each rule to provide a measure of categorisation accuracy. Once this was completed, participants were debriefed as to the aims and expectations of the study.

4.2.2 Results

Data from three participants was excluded for failing to meet the training accuracy criterion within the 45 minute session, leaving data from 81 participants for analysis. The data was first examined to determine whether categorisation accuracy was higher for the cued rule compared to the control, as was predicted by the Replay model. Figure 4.3 shows the average accuracy rates for both conditions. Accuracy was slightly higher for the uncued rule (M = 0.65 $\pm$ 0.02 95% confidence intervals) compared to the cued rule (M = 0.64 $\pm$ 0.02), though this was not a significant difference, paired $t(80) = 0.47$, $p = .640$, $d = 0.07$, suggesting no benefit for the cued rule.

The results also demonstrate that performance for both rules fell between training and test, with mean accuracy in both conditions falling below the training criterion of 0.75. Time away from the task therefore appears to harm performance whether experiences are cued or not, though cueing could perhaps help to offset this

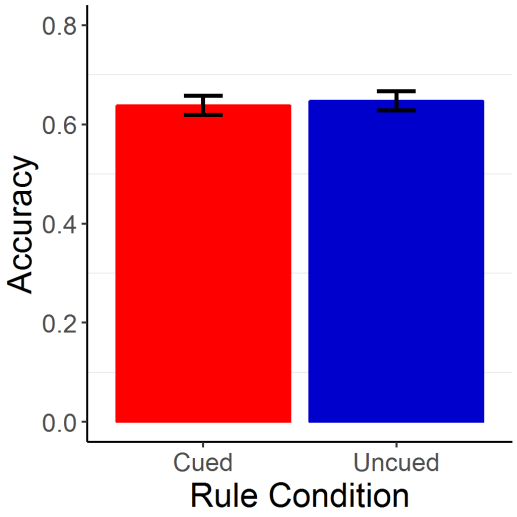

Figure 4.3: Average categorisation accuracy for the cued and uncued rules in the test phase of Experiment 1. Error bars show 95% CIs.

loss if performance started at a higher point. To compare this fall between the two rules, each participant's average accuracy rate across the final 16 trials of training for each rule was contrasted with test accuracy to measure the drop in performance. This drop was higher for the cued rule (M = 0.11 $\pm$ 0.02) than the uncued rule (M = 0.09 $\pm$ 0.02), though again this was not a significant difference, $t(80) = 0.95$, $p = .344$, $d = 0.15$. The fall in performance therefore appears to be fairly equivalent whether the rule was reinforced or not.

4.2.3 Discussion

Results from Experiment 1 do not support either of the hypotheses taken from the Replay model: accuracy rates for the cued categorisation rule were no better than either pre-break performance for that rule or test performance on the uncued control rule. As such, these results provide no evidence of a replay benefit in this task, with neither rule appearing to benefit from the break in training, whether cued or uncued. Instead, this disruption in training appears to lead to a fall in performance for both rules, suggesting the break led only to forgetting of what had been previously learned. The experiment therefore provides no support for our suggestion that rehearsal processes could permit the discovery of previously unknown rules, and so is

unable to demonstrate an association between replay processes and insight effects.

Three main explanations can be offered for such a result: first, that no replays were performed during the break; second, that the contextual manipulation failed to prioritise one set of trials over the other; and third, that the categorisation task does not benefit from trial replays.

The first of these explanations suggests a lack of any replays during the break period, thereby providing no rehearsal of either categorisation rule. This suggestion is supported by the drop in performance for both rules between training and test, implying that neither rule received any consolidation. This could be attributable to a lack of sufficient motivation to perform replays during the break, as participants gained no direct benefit for improved performance in the subsequent test period, providing little reason for consolidation. This is further compounded by the fact that participants had already met a performance criterion to advance from training to break, potentially suggesting performance was already at a satisfactory level and improvement was unneeded. While no measure of motivation is provided by this task, this is an important factor to consider in the study of rehearsal processes: it is uncertain whether learners will spend cognitive resources on consolidation without sufficient reason to do so, particularly where rest carries its own intrinsic reward (Kool & Botvinick, 2014, 2018). It is therefore advisable to better encourage the performance of replays when examining the potential benefits of rehearsal, for example by providing actual monetary rewards for accurate responses. Alternatively, cognitive resources needed for replays may have been directed elsewhere during the break, preventing any consolidation processes. While the break did include a vigilance task to maintain attention on the screen, this task was selected as the associated cognitive demands were not judged to be sufficient to disrupt trial replays. Even so, this could be verified by introducing a task/no task manipulation to the experiment, though the inclusion of an extra factor would likely require a larger sample than was used here.

The second explanation for these results suggests a failure of the contextual manipulation to prioritise one set of trials over the other, leading both rules to be re-

hearsed equally, and therefore generating no difference in performance between the rules in the test phase. This may be because context was not bound closely enough to the training trials to be used as a cue for replay, or because context is not in fact used to select trials for replay at all. While this would explain the lack of a performance difference between the cued and uncued rules, this does not account for the drop in performance observed between training and test; even if both rules were equally consolidated, this should have led to an increase in performance for both following the break. In addition, as previously noted, context has been successfully used as a cue in studies of memory consolidation (Rudoy et al., 2009; Bendor & Wilson, 2012), providing some support for its role in rehearsal processes. Failure of the contextual manipulation alone therefore does not seem to adequately explain the present findings.

The third and final explanation for these findings suggests that the categorisation task does not benefit from trial replays, meaning that neither rule was consolidated despite task manipulations being effective in encouraging rehearsal processes. This could occur if replays did not successfully isolate prediction errors in previous trials, leading to no correction of mistaken classifications, or if rehearsal continued to make use of overly simplistic, one-dimensional rules which were accurate in most, but not all, cases. Even so, this again does not account for the drop in performance in the test phase, as even one-dimensional rules would provide a higher accuracy rate than was observed for either rule, being correct in six of the eight possible items.

While the current data makes it difficult to definitively separate these explanations, the drop in performance following the break does suggest that the present results are due to a lack of rehearsal of either categorisation rule, whether cued or uncued. This provides two main directions for the present study: continue to use the difficult categorisation task, making slight alterations to examine or control for the factors noted above, or switch to alternative tasks which have previously demonstrated insight effects to determine whether these tasks might benefit from mental replays. Given the potential concerns as to the effectiveness of replay on the current

task, such a switch in method may be more valuable, particularly as this would allow for a closer connection between replay and insight. The second experiment in this chapter therefore continued to examine whether rehearsal processes could lead to the discovery of previously undiscovered solutions by switching to a novel task more closely connected to such discovery: anagrams.

4.3 Experiment 2: Anagrams

Anagram tasks are commonly used in the existing insight literature because solutions to these problems are often found through sudden, pop-out realisations after time away from the task without any apparent conscious deliberation, a key aspect of the insight phenomenon (Bowden, 1997; Novick & Sherman, 2003). However, while the solutions to these problems are often subjectively viewed as sudden and without consideration, there is also evidence to suggest that such solutions are in fact derived from the iterative testing of letter combinations during the incubation period, possibly indicating that a gradual process of unconscious mental testing underlies this realisation (Novick & Sherman, 2003; Penney, Godsell, Scott, & Balsom, 2004). In such cases, anagrams could then be solved using a method similar to that of the Replay model, with unsolved anagrams being revisited during periods of rest in a secondary attempt to find the solution.

This is admittedly not a direct correspondence; anagrams are unlikely to be solved using the same associative learning systems taken from the Rescorla-Wagner model. Even so, this process could follow a similar concept of allowing the learner more opportunities to attempt to solve the problem, so increasing the probability of identifying the true solution. Alternatively, revisiting past failures could help to solidify which responses are incorrect, assisting the direction of subsequent solution attempts. Such a suggestion is supported by studies finding solution rates to be higher following longer incubation periods (C. Peterson, 1974; Goldman, Wolters, & Winograd, 1992), though whether this is due to greater consideration of the problem is still under debate (e.g. Vul & Pashler, 2007).

As such, anagrams offer a good crossover between existing consolidation models and theories of insight, making them useful in further examinations of the link between insight and mental rehearsal. Experiment 2 therefore sought to apply an anagram solving task to the consolidation design of Experiment 1: by encouraging the replay of unsolved anagrams in a period of rest, performance on those anagrams may be improved in a later test period. This again made use of a contextual manipulation both to encourage the performance of replays as well as to provide a contrast between cued and uncued trials: by associating anagrams with particular contexts, placing the learner back in a context should prioritise replay of the missed trials from that context, thereby rehearsing those errors more than other stored trials. If such rehearsal does indeed assist in finding the solution to the replayed trial, then this should lead to a higher probability of solution for unsolved anagrams from the re-shown context compared to the control context in a subsequent test. Alternatively, if solutions are equally likely between contexts in this period, then this rehearsal may instead lead to an advantage in solution speed, with errors from the re-shown context being solved faster.

4.3.1 Method

The hypotheses, experimental design and planned data analysis for Experiment 2 were preregistered on the Open Science Framework before data collection began. Full details can be found at: https://osf.io/svc7j/.

Participants

One-hundred-and-twenty-six participants were recruited from the University of Warwick online SONA system in return for financial compensation, made up of an initial payment of £2, and a performance bonus ranging between £0 and £6. The sample included 86 females and 39 males (1 declined to answer), while age ranged between 18 and 30 years, with a mean of 20.7.

Design and Materials

Experiment 2 used an anagram solving task, in which a 5-letter single-solution word scrambled into a randomised order appeared on-screen for a set period of time, and participants were asked to enter the correct word using the keyboard (for example, the scramble BMALU may appear, with the solution being ALBUM). Trials ended when a guess had been made or when the time limit was met, at which point the scramble was immediately removed from the screen, and participants were informed of their accuracy. Response accuracy and reaction time were then recorded before advancing to the next trial.

Anagrams were drawn from a list of 204 5-letter single solution words taken from Gilhooly (1978). These words were randomly scrambled before the experiment began such that each participant viewed the same order of scrambled letters for a given word.

As in Experiment 1, trial context was composed of a background image of a location and a coloured rim, with each being randomly selected at the start of the experiment to form two distinct compound contexts (illustrated in Figure 4.4).

Experiment 2 also included financial rewards for performance in order to address the potential motivation issues in the previous experiment noted above. Correct answers were therefore awarded varying numbers of points, with the final point total accumulated across the task being converted into a bonus payment at the end of the experiment.

Figure 4.4: An example slide from the training phase of Experiment 2.

Procedure

Upon arriving at the lab, participants were first provided with a short, written description of the experiment to introduce the task. As the task requires familiarity with English words, participants were also asked to describe their fluency in English as either poor, satisfactory, good or excellent (a non-disclosure option was also available).

The experiment was divided into multiple rounds, each divided into 3 phases: training, break and test. In each round, participants began by solving anagrams until meeting an error criterion, followed by a break to allow for rehearsal, before being given a second opportunity to solve the failed anagrams. Multiple rounds were used to avoid overloading participants with a high number of failed trials to replay during the break; results from each round were then aggregated to provide a sufficient number of measurements for analysis.

During the training phase, participants performed anagram trials in each of the two distinct contexts. Trials were solved in blocks of five in one context before switching to the other context. Initially, participants were given 10 seconds to solve each anagram, though this decreased by 2 seconds with each completed block of five trials to a minimum of 2 seconds. This decrease was intended to control for ceiling performers by increasing difficulty over the course of the training phase. Points were awarded for correct solutions, starting at 1 point for the first correct answer in a context and increasing by 1 point with each successive solution up to a maximum of 10 points per correct answer in each context. This point scheme was intended to encourage participants to attempt to remain in the training phase for as long as possible.

After making two errors in a given context, participants no longer viewed trials in that context; the training phase therefore ended once two errors were made in each of the two contexts. Before moving to the break phase, participants were told that they would be given a second chance to solve the anagrams they missed, but only after a 1-minute interval. Participants were also told that solutions in this

second chance period would be awarded double the original point value of that trial; this was intended to encourage rehearsal of the failed trials during the break phase. Answering correctly during the training phase would still however provide access to higher-value anagrams, so limiting the potential strategy of deliberately answering incorrectly for known solutions in order to obtain doubled rewards in the later test phase. A slide was shown at this point to remind participants of the scrambles that had been answered incorrectly to control for recency effects in potential consolidation.

The break phase then began, in which one context was randomly selected from the preceding training phase and re-shown for one minute. Break length was reduced in this task due to the use of multiple breaks across rounds, though the number of trials available for replay was also lower than in the previous experiment. During this time, participants were asked to perform a vigilance task to maintain attention on the screen: Xs appeared on-screen in random locations at 3- to 6-second intervals during the break, and participants were asked to click the Xs as close to the centre as possible.

Once the break was complete, participants moved to the test phase, in which the 2 errors from each context were re-shown and participants were given a second opportunity to attempt solution. As previously noted, correct responses in this phase were awarded double the point value of the original trial. The solution (correct/incorrect) and reaction time from each trial were then recorded for later comparison between the cued and uncued contexts.

Completing the test phase marked the end of that round, and a new round began. Rounds continued until all anagrams were viewed, or until a maximum of 10 rounds were completed. Once either of these criteria was met, the experiment ended, and participants were debriefed as to the aims and expectations of the study. The total number of points earned by the participant was then translated into a bonus payment, with each point being worth £0.01.

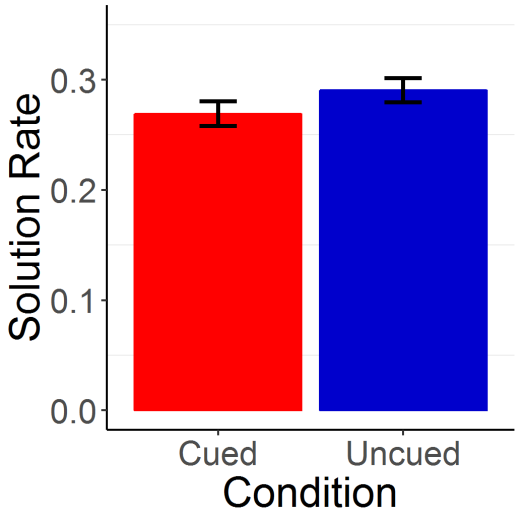

Figure 4.5: Average solution rates from the cued and uncued trials in the test phase of Experiment 2. Error bars show 95% CIs.

4.3.2 Results

Participants were excluded from the analysis if the participant failed to finish the experiment or if the participant did not make at least 12 total errors across the experiment to allow for later comparisons, though no participants met these criteria.

Performance in the task was found to be fairly low in both the training and test phases: average solution rates were 0.33 ($\pm$ 0.03 95% confidence intervals) in training and 0.28 ($\pm$ 0.02) at test, suggesting the task was reasonably difficult. To assess the impact of break context on performance, contrasts were first made between average solution rates across test trials from the cued and uncued contexts, shown in Figure 4.5. Rates were higher by 0.02 ($\pm$ 0.02) in the uncued context (M = 0.29 $\pm$ 0.01) compared to the cued context (M = 0.27 $\pm$ 0.01), though this difference did not meet the standard significance threshold, paired $t(125) = 1.91$, $p = .058$, $d = 0.16$. This result therefore opposes the above hypothesis, with missed anagrams from the cued context being no more likely to be solved than those from the control context. In fact, this demonstrates a near significant difference in the opposite direction, suggesting that this cueing may have if anything harmed later performance. Even so, this difference does not meet the criterion for statistical significance, suggesting this may not be a reliable effect.

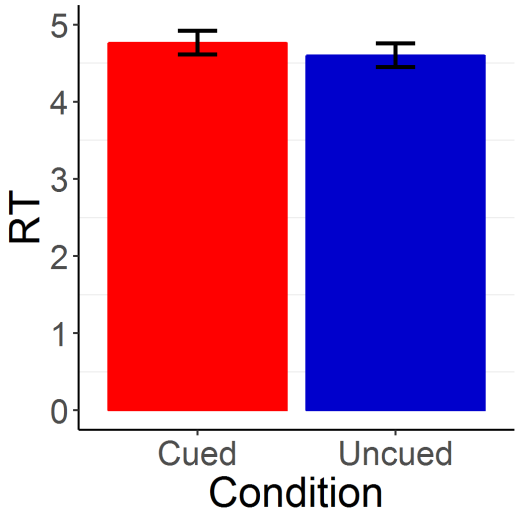

Figure 4.6: Average response times from the cued and uncued trials in the test phase of Experiment 2. Error bars show 95% CIs.

As a follow-up to this test, contrasts were also made between the average response times from the two conditions, shown in Figure 4.6. Anagrams were solved 0.16 ($\pm$ 0.30) seconds faster in the uncued context (M = 4.60 $\pm$ 0.15) compared to the cued context (M = 4.76 $\pm$ 0.15), though this difference was again non-significant, paired $t(125) = 1.07$, $p = .287$, $d = 0.12$. Cued errors were therefore not solved reliably faster than those from the control context, though this contrast is less meaningful given the generally low level of performance at test, as a difference in solution time is only relevant where anagrams were being solved frequently in both contexts.

As a secondary analysis, we next examined the relationship between the frequency of training trials and later test performance in order to test for any potential interference in selecting failed trials for replay: due to the use of an error criterion in training, the frequency of training trials performed by each participant was variable, with high-performing participants viewing more trials than low-performing participants. This difference in training frequency may have inadvertently introduced an additional factor to the rehearsal process: higher training frequencies provide a larger sample of trials available for replay, potentially interfering with the rehearsal of the key failed trials examined in the test phase by making the selection of those

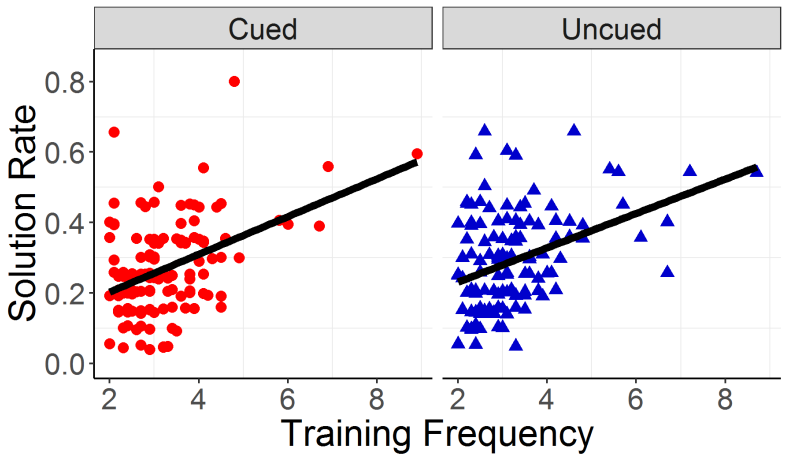

Figure 4.7: Comparisons of training frequency and test solution rate from the cued and uncued contexts of Experiment 2, including lines of regression.

trials less probable.

Figure 4.7 shows average training frequencies against test solution rates from the two conditions, demonstrating that participants who received more training trials tended to perform better in the test phase. This was assessed using correlations between training frequency and solution rate, which were significantly positive in both the cued ($r = 0.424$, $t(124) = 5.22$, $p < .001$) and uncued contexts ($r = 0.390$, $t(124) = 4.72$, $p < .001$). As such, rather than any interference due to higher levels of training, the results instead indicate that participants who performed well during training and therefore viewed more trials also performed better at test, likely reflecting underlying skill at the task.

4.3.3 Discussion

The above results provide no evidence of the hypothesised replay benefit in Experiment 2, with anagrams from the re-shown context being no more likely to be solved following the break than the uncued trials. The experiment is therefore unable to support the suggestion that insight-like effects often observed in anagram solution are due to the replay of unsolved problems between solution attempts assisting the learner in identifying the correct answer.

The potential reasons for this result are broadly the same as those offered

for Experiment 1: replays may not have been performed in the task at all, replays may not have been adequately biased by trial context, or the anagram task may not have benefited from the replay of failed trials. This is despite the adjustments made to the design of Experiment 2 that attempted to address some of these concerns: financial rewards for high performance were intended to better encourage the replay of missed trials, while the switch to the anagram solving task was intended to account for possible issues in the categorisation task by moving to a design that had previously demonstrated insight effects.

It is notable, however, that while the anagram task is more closely aligned with insight, this is also less connected to the associative learning tasks on which the Replay model was built, a potential cause for concern when applying the principles of this model to the task. This application was based on the assumption that providing the learner with more opportunities to revisit past failures through trial replays would increase the probability of solution, either by allowing more solution attempts or by identifying the reasons for previous mistakes; such a prediction builds on the findings of past studies suggesting insight effects in anagram tasks are the result of a process of incremental learning across the incubation period (e.g. Novick & Sherman, 2003; Penney et al., 2004). Instead, the results appear to indicate that prompting such attempts may if anything harm later performance, though the reliability of this effect remains uncertain. This relates to the concept of fixation within the insight literature where learners are suggested to become fixated on incorrect responses, preventing the realisation of the true solution. In such cases, the incubation period is suggested to allow the learner to move away from these errors, so starting afresh on any subsequent solution attempts (Vul & Pashler, 2007).

This could then explain the present results: cueing the replay of failed trials could lead to a fixation on previous mistakes, blocking further attempts to find novel paths to solution. If so, this could mean that anagram tasks are unable to benefit from the consolidation of learning through replay at all, being based in a different problem-solving system separate from the associative learning processes underlying the Replay model. This connects with the wider debate on the nature

of incubation effects in anagram tasks noted above: studies in the insight literature have argued both that such effects are based on an incremental accrual of evidence in this period (Novick & Sherman, 2003; Penney et al., 2004), as well as a true insight process requiring the abandonment of previous failures (Vul & Pashler, 2007). The findings of this experiment are unfortunately unable to provide a substantial contribution to this debate given the lack of a significant effect in the observed results, simply displaying a scenario in which contextual cueing did not appear to reliably affect the solution rate of related anagrams.

These results do, however, provide some guidance in the continuation of the present research into the role of learning consolidation processes in the discovery of previously unnoticed concepts: if trial replay is to be applied as a potential mechanism to allow for such discovery, this requires some assurance that this form of consolidation will be effective in generating discovery, including greater clarity on the way in which this process alters the existing mental representation. It is therefore advisable to return to the use of paradigms with a greater correspondence to the principles of the Replay model where the effect of trial replay is better understood, allowing for more definitive model predictions: associative learning tasks. Insight tasks, meanwhile, may need to be examined on a case-by-case basis to determine whether consolidation processes could be applicable as underlying mechanisms.

It is also notable that this provides a second occasion in which context may not have effectively biased replay selection. While there are in both cases results which point towards other reasons for this failure (the performance drop in Experiment 1 and the possible fixation in Experiment 2), this could also indicate that this is in fact an ineffective manipulation, with context either not being used as a selection criteria, or requiring a closer connection to trial events to be used in this way. As such, it may also be advisable to move away from the use of a contextual manipulation as a precaution for such issues.

The third experiment in this section therefore made a greater departure from the design of the previous tasks, both selecting a more traditional associative learning design to facilitate the application of the Replay model, as well as replacing

the potentially faulty context manipulation. This involved the consideration of a number of existing paradigms in order to identify a design more closely related to associative learning principles in which learning consolidation processes could lead to the discovery of previously unnoticed concepts; this ultimately led to the selection of a sensory preconditioning task, being both an established associative learning effect as well as a display of the apparent integration of separate sets of learning into new behaviours.

4.4 Experiment 3: Sensory Preconditioning

Sensory preconditioning refers to an effect in which a value or response trained to one stimulus is seemingly passed to an untrained but associated stimulus; for example, in the original description of the effect by Brogden (1939), following training on a light/bell compound, dogs trained on a shock response to the bell later showed a similar response to the light, despite never having seen the light paired with a shock. In its most basic form, this can be divided into three phases, as illustrated in Figure 4.8: in Phase 1, learners are trained on associations between two sets of stimuli (A+B); in Phase 2, learners are trained on a value or response for one of these sets (B+R); and in Phase 3, learners are tested on their response to the untrained stimulus (A) to examine any transfer. The present study focuses on the value case, in which the trained stimulus is rewarded, and Phase 3 tests the transfer of this reward to the previously untrained stimulus; for instance, the learner may be given choices between stimuli which were and were not paired with the trained item to assess which is preferred (e.g. Wimmer & Shohamy, 2012). This procedure

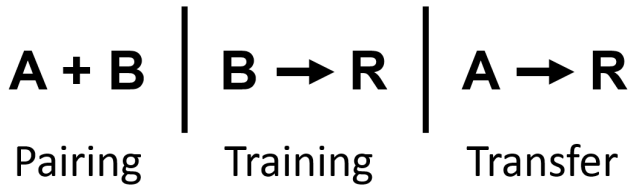

Figure 4.8: Illustration of the basic three-phase sensory preconditioning design.

has been observed to demonstrate such a value transfer, suggesting internal mental processes may propagate the value of one item to another despite never actually seeing the untrained item lead to reward (e.g. J. L. Jones et al., 2012; Wimmer & Shohamy, 2012).

Sensory preconditioning therefore appears to demonstrate the combination of separate sets of learning to direct future choice without direct training, making it well suited to the present examination of learning consolidation. An additional advantage of sensory preconditioning in the present study is that it can be explained using associative learning systems: due to the training of an association between the rewarded and unrewarded stimuli in Phase 1, when the rewarded stimulus is presented in Phase 2, the memory of the unrewarded stimulus is also activated, leading both stimuli to be connected with the subsequent reward (Wimmer & Shohamy, 2012). This basis in associative theory then allows for easier application of the Replay model to this process: if transfer results from such memory activation, then additional replays of Phase 2 trials should then amplify this process, leading to a stronger effect. In this case, replay does not generate the discovery of broader associative structures, but could strengthen existing effects, making such discovery more apparent. Providing greater opportunity for replays could then further facilitate the preconditioning effect, solidifying the transfer of value from one stimulus to another. This can again be achieved by adding a break between training and test, offering the learner more time to perform replays and thereby increasing the level of consolidation.

The third experiment of this section therefore uses a sensory preconditioning design to contrast the transfer of value from trained to untrained stimuli between participants who are given a break to perform replays after training with those without such an opportunity. This replaces the contextual manipulation of the previous experiments with a simple break/no-break comparison, removing the previously stated concerns regarding the replay selection criteria. If trial replay amplifies the sensory preconditioning effect, then participants in the break condition should show greater transfer than those receiving no break.

The application of replay to this paradigm does however allow for the rehearsal of trials from both Phases 1 and 2, potentially leading to additional effects: while replay of Phase 2 trials could amplify transfer as suggested above, replay of Phase 1 trials after learning reward associations could also lead the learner to infer that the untrained stimulus was obstructing reward in the first phase. This would then cause the untrained stimulus to be treated as an inhibitor to reward, negating the value of the rewarded stimulus. This potential outcome is similar to the backwards conditioned inhibition effect (Chapman, 1991; Urcelay, Perelmuter, & Miller, 2008), in which the training of a paired stimulus without reward (XY-) followed by rewarded training with one of those stimuli (Y+) leads the unrewarded stimuli (X) to become inhibitory. Such a procedure bears a strong resemblance to the described sensory preconditioning effect, but with opposing results: rather than gaining the value of the rewarded stimulus, the untrained stimulus instead gains the negative of this value in order to counteract this reward. This effect corresponds with the associative framework of the Replay model: when replaying trials from Phase 1 after learning the reward values in Phase 2, the presence of the trained stimulus leads to the expectation of reward for these trials. Given that no such reward is given in these Phase 1 trials, the associative strength between both presented stimuli and reward decreases; in the case of the trained stimulus, this is offset by replays of the rewarded Phase 2 trials, maintaining the association, whereas the association of the untrained stimulus falls below its initial starting point of zero, becoming an inhibitor.

The Replay model is therefore able to predict two opposing results from the same consolidation process: one in which the untrained stimulus continues to gain the value of its trained associate, and one in which it gains the negative of this value. This results from potential differences in the way in which trials from the first phase are represented when selected for replay: these trials may be viewed as unrewarded in contrast with the Phase 2 trials, leading to the inference of inhibition described above, or they may simply be considered as being uninvolved with reward information, showing only the co-occurrence of the two presented stimuli.

The demonstration of either standard preconditioning or inhibition following the break period could then constrain the representation used by the Replay model, suggesting the inferences made by the learner when identifying broader learning patterns during consolidation.

Experiment 3 therefore uses an adapted form of the described preconditioning design in which a break/no-break condition is added between training and test, providing a contrast in the degree of value transfer from trained to untrained stimuli between scenarios offering either high or low opportunity for replay. We examine two hypotheses regarding the effect of this manipulation based upon the predictions of the Replay model described above:

Hypothesis 1: Amplified Preconditioning

If replays of Phase 2 trials amplify existing preconditioning effects, then choices between untrained stimuli in the test phase should demonstrate greater correspondence with the value of their trained associates following the break, leading to a greater preference for untrained stimuli with high-value associates in the break condition.

Hypothesis 2: Inhibition

If replays of Phase 1 trials lead to the inference that the untrained stimuli obstructed rewards in that phase, then these items should gain the inverse value of their associates; as such, choices at test should then prefer untrained stimuli with lower-value associates in the break condition under the assumption that these are less inhibitory.

4.4.1 Method

As with Experiment 2, the hypotheses, experimental design and planned data analysis for Experiment 3 were preregistered on the Open Science Framework before data collection began. Full details can be found at: https://osf.io/venbp/.

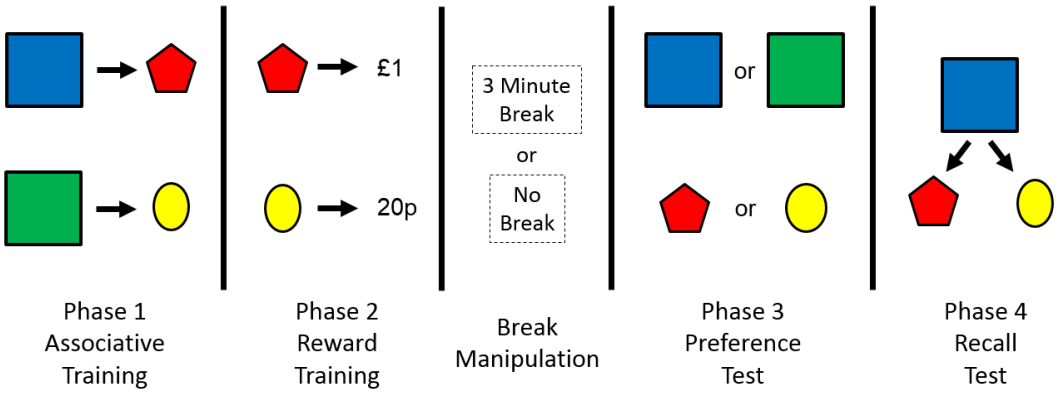

Figure 4.9: Illustration of the design of Experiment 3, including two example stimuli from the Untrained (boxes) and Trained (shapes) sets.

Participants

Ninety-nine participants were recruited from the University of Warwick online SONA system in return for financial compensation, made up of a base payment of £2 plus a bonus of £0-£3 dependent on performance in the test phases. The sample included 57 females and 32 males, while age ranged between 18 and 41, with a mean of 22.2.

Design and Materials

The experiment used an adapted form of the three-phase sensory preconditioning design described above, with two main alterations: first, a break/no-break manipulation was added to the task, with participants in one condition receiving a three-minute break between Phases 2 and 3; second, a recall test was added to the task as a fourth phase both to assess whether stimulus associations were remembered as well as to distract participants from the importance of Phase 3 trials as a measure of transfer. The task was therefore divided into four phases: associative training, reward training, preference test and recall test, illustrated in Figure 4.9.

Six stimulus pairs were used in the experiment, each made up of a 'Trained' stimulus and an 'Untrained' stimulus drawn from respective sets; these were represented as coloured boxes (Untrained set) containing geometric shapes (Trained set), with the shapes being worth different amounts of money. Colours and shapes

were randomly assigned to sequences at the start of each run of the experiment, and remained fixed and deterministic for each participant. Monetary values were used to provide a reward scheme for performance in the test phases of the task, with participant responses being directly tied to actual monetary gains according to their choice. This scheme used a binary high/low format for the Trained items, with three shapes having high values (£1) and three having low values (£0.20); these values received variation in the task by adding random noise figures, drawn on each trial from a uniform distribution between -£0.15 and £0.15. Rewards were explicitly based upon choices in Phases 3 and 4, thereby encouraging accurate learning of the sequences.

Procedure

Upon arriving at the lab, participants were first randomly assigned to either the break or the no-break condition; this was balanced to provide approximately equal numbers in each group, meaning 45 participants were assigned to the break condition and 44 participants were assigned to the no-break condition. Participants were told the experiment examined the learning of stimulus sequences, and would involve learning both simple sequences of items as well as the value of those items, with a subsequent test to assess performance.

The task began with Phase 1, in which a single stimulus from the Untrained set was presented on screen, and participants were asked to press a key on the keyboard to see which stimulus from the Trained set followed. After a response, the associated Trained stimulus appeared on-screen for 1 second, before advancing to the next trial. Each Untrained stimulus was presented five times across the phase for a total of 30 trials, with the order of stimuli being randomised.

After completing Phase 1, participants moved to Phase 2, which followed a similar structure: a stimulus from the Trained set appeared on-screen, and participants made a key press to see the monetary value of that item. Values were generated on each trial according to the fixed mean value of the displayed Trained stimulus plus a randomly sampled noise term, described above. Again, each Trained

stimulus was presented five times across Phase 2 in a randomised order for a total of 30 trials.

At this point, participants in the break condition moved to an alternate task for 3 minutes to allow for the performance of replays; this was prefaced by a slide reminding participants that the break would be followed by a test of their memory for the preceding sequences, intended to encourage rehearsal. Break length was again set at 3 minutes based on the design of Gershman et al. (2014), matching with Experiment 1. As with the previous experiments, the break included a vigilance task to maintain attention on the screen, in which participants were asked to click Xs as they appeared on-screen at 3-6 second intervals. Participants in the no-break condition moved directly to the next phase.

Participants then began Phase 3 of the task, in which two stimuli from the same set were presented on either side of the screen, and participants were asked to choose which they would prefer based on their previous experiences in the earlier phases. Choices were made using one of two corresponding keystrokes representing either the left or right stimulus, and led to reward values corresponding with the trained sequence, meaning selection of an Untrained stimulus that led to a high-value Trained stimulus was treated as a high-value choice. Participants did not, however, receive rewards during this phase, though choices between high- and low-value stimuli were recorded and used to generate bonus payments at the end of the task. As such, participants received no feedback on the outcome of their selections, preventing any contamination of future decisions by that feedback.

Participants completed all 15 potential comparisons from each stimulus set twice, plus four additional repetitions of the 9 key comparison trials (high vs. low) from the Untrained set, for a total of 96 test trials, again presented in a randomised order. While choices between the Untrained stimuli provided the main measure of the experiment, choices between the Trained set offered a verification that participants had accurately learned the value of these items in Phase 2.

Finally, Phase 4 provided participants with a recall test, in which an Untrained stimulus appeared at the top of the screen, and two Trained stimuli (one

correct, one randomly selected foil) appeared on the left and right. Participants were instructed to select which of the two Trained stimuli followed after the shown Untrained stimulus, again responding using keystrokes to select either the left or right shape. As with choices in Phase 3, participants did not receive feedback on their answers to prevent contamination of future responses, though these were again used to determine the final bonus payment. As with Phases 1 and 2, each Untrained stimulus was presented five times across Phase 4 in a randomised order, for a total of 30 trials.

After completing all four phases, the experiment ended, and participants were debriefed as to the aims and expectations of the study. Bonus payments were then calculated by randomly selecting a key comparison trial from Phase 3 and a recall trial from Phase 4, and using the participant's responses in these trials to determine a reward payment. Choice trials were rewarded following the same scheme as Phase 2, providing either a high or low value plus noise, while recall trials received a fixed reward of £1 if correct.

4.4.2 Results

Figure 4.10 shows average preference rates for the high-value Untrained items across the key high/low comparison trials from Phase 3. Average rates were marginally higher in the no-break condition (M = 0.63 $\pm$ 0.08 95% confidence intervals) than the break condition (M = 0.61 $\pm$ 0.08), though this was not a significant difference, $t(97) = 0.25$, $p = .801$, $d = 0.05$, indicating no reliable effect of break on preference. This rate was, however, found to be significantly higher than 0.5 in both the break ($t(44) = 3.05$, $p = .004$, $d = 0.45$) and no-break ($t(43) = 3.32$, $p = .002$, $d = 0.50$) conditions, suggesting participants were not answering randomly in these trials.

As a secondary test, preference rates were also compared across the high/low comparisons from the Trained set to check learning of stimulus values, shown in Figure 4.11: this again was higher in the no-break condition (M = 0.785 $\pm$ 0.075) than the break condition (M = 0.749 $\pm$ 0.067), though this was not a significant

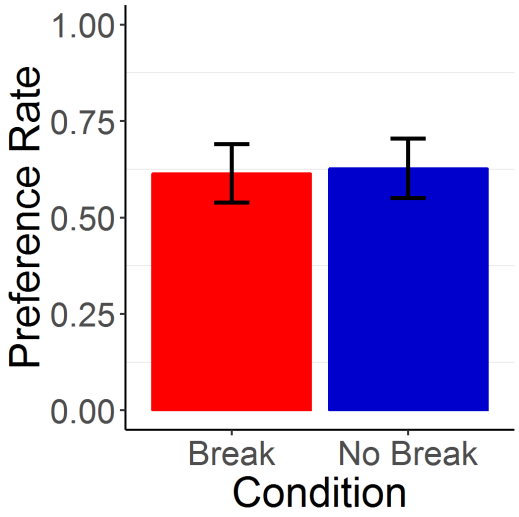

Figure 4.10: Average preference rates for high-value Untrained items in the test phase for the break and no-break conditions of Experiment 3. Error bars show 95% CIs.

difference, $t(97) = 0.72$, $p = .471$, $d = 0.16$ again suggesting no reliable effect of break. This was again however found to significantly differ from chance selection in both the break ($t(44) = 17.5$, $p < .001$, $d = 1.12$) and no-break ($t(43) = 16.7$, $p < .001$, $d = 1.16$) conditions, demonstrating reasonably accurate learning of shape values in both groups.

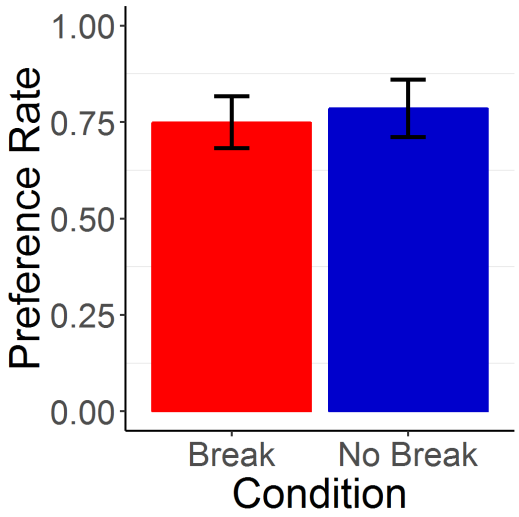

Figure 4.11: Average preference rates for high-value Trained items in the test phase for the break and no-break conditions of Experiment 3. Error bars show 95% CIs.

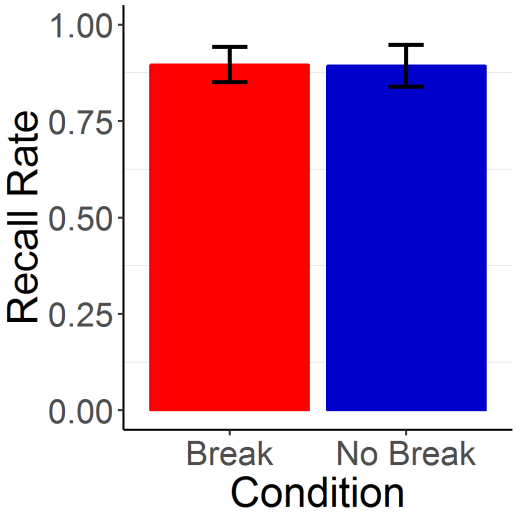

Figure 4.12: Average recall rates for stimulus associations in the recall phase for the break and no-break conditions of Experiment 3. Error bars show 95% CIs.

Similarly, recall rates were also compared between groups to assess memory for the initial associations, shown in Figure 4.12: these rates were higher in the break condition (M = 0.896 $\pm$ 0.045) than the no-break condition (M = 0.893 $\pm$ 0.054), though this was once again a non-significant difference, $t(97) = 0.09$, $p = .929$, $d = 0.02$, indicating no reliable effect of break on memory for the sequences. As with the previous measures, however, recall rates were above-chance accuracy in both the break ($t(44) = 17.6$, $p < .001$, $d = 2.63$) and no-break ($t(43) = 14.8$, $p < .001$, $d = 2.22$) conditions, indicating these associations were accurately acquired in both groups.

Finally, correlations between preference rates from the Trained and Untrained sets were used to assess the relationship between learning and transfer; this was found to be positive, $r = 0.663$, $p < 0.001$, suggesting that better learning of the value of Trained stimuli led to greater transfer to the Untrained stimuli.

4.4.3 Discussion

The findings of Experiment 3 appear to conform with previous displays of sensory preconditioning, with participants showing a greater preference for items with high-value associates, implying that these items were viewed as holding similarly high

values. This reflects a transfer of value from trained to untrained stimuli based on internal mental processes rather than any new observations: participants appear to be inferring the value of the untrained items via their relationship with their rewarded associates, so using the combination of distinct sets of learning to guide their behaviour in the choice trials.

Preferences did not, however, differ with a break between training and test, suggesting that such transfer is unaffected by the amount of time available for rehearsal, so opposing both the amplification and inhibition hypotheses taken from the Replay model. Such results can again be attributed to the explanations offered for the lack of replay effects in the previous tasks: while the removal of the contextual manipulation eliminates potential failures in replay bias, the lack of a break effect again implies either that participants may not have performed replays across the break period, or that replays may not have led to any substantial change to the participants' associative representations.

In the case of the former explanation, this again reflects concerns regarding participant motivation or distraction interfering with rehearsal despite the continued use of design elements to counteract these issues. There is however an additional factor relating to this explanation introduced in this task due to the removal of the contextual manipulation: without this element, the present design provides no direct cue towards replay in the break period at all, offering less assurance of rehearsal. While the contextual manipulation did not appear effective in the previous experiments, the lack of any method of encouragement in this task remains a notable issue, and should be addressed in future studies of replay. It is therefore advisable to determine alternate methods of prompting rehearsal to more adequately assess this conclusion; for example, context could be more closely bound to training trials to assure its use as a cue, or the training stimuli themselves could be re-presented during the break period, though this could introduce other confounds.

The latter of these explanations meanwhile reflects deeper concerns that replays may have little influence on sensory preconditioning tasks even where rehearsal is effectively encouraged. This would be a more surprising result for this

task given that sensory preconditioning was specifically chosen in this experiment for its existing basis in the same associative learning principles that form the basis of the Replay model: Replay uses the same associative learning framework which seemingly underlies both the acquisition of stimulus and reward associations as well as the transfer of reward values displayed in this task. It is therefore unclear why the same process which accounts for these behaviours during training should fail to generate change in offline rehearsal if replays are occurring as described.

Such a finding could then relate to other potential issues introduced when adding a rehearsal period to the present design beyond learning consolidation. A key example of such a factor is memory decay: longer intermissions between training and test could also cause greater losses in trial memory, counteracting any potential rehearsal. Adding a break period therefore introduces a conflict between rehearsal and forgetting, an aspect which is more pronounced in the current design due to the contrast between break and no-break conditions rather than the contextual manipulations of Experiments 1 and 2. If so, this would have to be a fairly equal balance in the present task given the lack of break effect in the above results, with forgetting essentially negating any rehearsal benefit to generate an equivalence between conditions. What is more, this would also call into question the global benefits of consolidation across tasks, being consistently counteracted by memory failures. This may then need to be assessed more directly to determine the impact of forgetting on rehearsal, for example using different break lengths to examine the time course of the effect.

There is also the possibility that the two opposing effects hypothesised by the Replay model did actually both occur during the break period, but due to their conflicting directions, ultimately cancelled each other out: if participants replayed trials from both Phases 1 and 2 during the break at approximately equal rates, then the initial preconditioning effect may have been both alternately amplified and counteracted, leading to no overall difference. This is however a highly speculative explanation based on the operations of the Replay model; the present data is unable to confirm either individual effect occurred in this task, and so cannot verify such

a suggestion. This could however be tested by cueing participants to replay either Phase 1 or Phase 2 trials to separate the two effects, though this would again require more reliable cueing mechanisms as noted above to effectively promote rehearsal of one set of trials over another.

The current data is unfortunately unable to effectively discriminate between these explanations, providing no clear indication as to whether replays were or were not performed in the task. These results do however indicate that potential rehearsal benefits are dependent on more than just greater opportunity for replay: simply providing learners with a break to perform replays does not guarantee a greater benefit, as other factors could interfere with rehearsal in this period. This then suggests that consolidation processes are more complex than was assumed by the initial depictions taken from the Replay model, involving the consideration of more factors beyond rehearsal time. As such, Replay may require further development to capture these elements if it is to provide more accurate predictions. This applies both to the theoretical side, expanding on model definitions to account for these factors, as well as the empirical side, adapting experimental contrasts to provide more complete assessments of model predictions.

While the lack of a break effect makes the processes involved in this task ambiguous, it does remain notable that in replicating sensory preconditioning effects, this data demonstrates the discovery of overarching structures outside of direct training that was initially targeted by this study; independent of the break manipulation, participants are combining separate sets of training to guide their later responses. The present sensory preconditioning task does therefore appear to offer a good base for further examinations of such discovery, even if the role of rehearsal in this task remain unclear. This may then be a valuable design for continued examination of consolidation results, offering a building point for future work to more closely investigate the systems involved in learning consolidation.

4.5 General Discussion

Across the three experiments of this section, we have found no definitive evidence that mental rehearsal of past experiences through trial replays assists in the discovery of previously unrealised data structures: Experiments 1 and 2 find no benefit from cueing such replays, with Experiment 1 in fact observing a drop in performance, while Experiment 3 finds no benefit from providing greater opportunity for replay using a break period. This then makes it difficult to suggest the value of replays in building more accurate representations of real-world data patterns, offering no clear evidence of learning consolidation processes having any impact on mental structures. Such findings also prevent any connection between general learning consolidation systems and insight effects, particularly given that none of the three tasks studied here demonstrate any improvement with time away from training.

These findings then provide an interesting contrast with the existing literature on rehearsal and consolidation in memory which provided the basis for the present work: despite the integral role of these processes in forging long-term memory (Atkinson & Shiffrin, 1968; McGaugh, 2000; Ratcliff, 1990), we here find no impact of similar systems on associative representations in three distinct tasks. This could then present a disconnect between the systems supporting learning and memory, with the former being less reliant on rehearsal of past experiences than the latter. Such a distinction would be somewhat surprising given the potential benefits of rehearsal in maintaining an accurate representation, as well as the fit of such a system to existing behaviour (Ludvig et al., 2017; Chapman, 1991; Ratcliff, 1990). This is in addition to the apparent display of such rehearsal in neural structures (Wilson & McNaughton, 1994; Euston et al., 2007; Davidson et al., 2009), as well as the attribution of existing associative learning effects to such processes (Gershman et al., 2014; Momennejad et al., 2018). The role of rehearsal within learning systems therefore appears somewhat ambiguous, perhaps suggesting a more complex interaction between direct learning and subsequent consolidation than was initially assumed in this study.

Despite these results, there are two potential suggestions that could be made regarding the role of replay based on common elements identified throughout these studies: first, replays appear to be difficult to effectively encourage; and second, replays may not assure a benefit in all scenarios. The first of these suggestions refers to the potential issues noted in all three experiments regarding the manipulations used both to prompt and bias trial replays: while Experiments 1 and 2 used trial context to provide a contrast between cued and uncued conditions, the lack of effect in these studies casts some doubt on the success of this design. Conversely, by removing this manipulation from Experiment 3, this task included no direct cueing mechanism at all, providing even less assurance that replays were being performed. This is a substantial cause for concern given that any assessment of the benefits of the rehearsal of past learning depends on reliable experimental contrasts between situations where learners are and are not performing replays, so requiring some assured method of encouragement. As such, it appears necessary to more definitively identify the criteria by which experiences are selected for replay in order to provide more concrete comparisons in future examinations of learning consolidation. In this respect, the present findings could be useful as indications that trial context is not used as such a criterion, demonstrating no effect in two different tasks. This is not, however, a definitive conclusion, as there are additional factors within each of these tasks that may have interfered with replay; the role of context may then need to be revisited in future work to be more conclusively eliminated, alongside other potential cues.

The second suggestion is based primarily on concerns raised in Experiments 1 and 2 that the tasks in question may not benefit from trial replay even where rehearsal occurs, though similar concerns could be raised in Experiment 3 given the lack of break effects. Such a suggestion also matches with the apparent distinction in findings between the present experiments and previous studies which indicated rehearsal effects in associative learning paradigms (e.g. Gershman et al., 2014; Chapman, 1991): these differences in results could be attributed to differences in the tasks used in these studies, with the current tasks offering poorer applications

of rehearsal than those used in the existing literature. This could then imply that Replay is a somewhat limited system in terms of generality, only able to generate benefits in a subset of tasks that correspond with the model's framework. Such issues again promote the careful consideration of the tasks used when examining the validity of Replay in order to maintain focus primarily on those likely to display rehearsal benefits. This was noted previously when discussing the results of Experiment 2, leading to the suggestion of restricting attention to tasks more closely aligned with the associative learning principles that form the foundation of the Replay model. It is therefore advisable to maintain this focus in future work, assessing the Replay model within its existing framework before attempting to adapt or extend the model to more varied tasks.

Such a suggestion also relates to concerns regarding memory decay noted in Experiment 3: even where rehearsal benefits could be gained from time away from a task, this may be counteracted by the loss of trial memories during the same period, again limiting the impact of replay on the representation. The collected results offer conflicting evidence on this interaction: while the drop in performance of Experiment 1 does suggest a loss of trained patterns, Experiment 3 finds no general differences in memory with a break, while Experiment 2 provides no real measure of memory (though this task did use a reminder slide to counteract such concerns). Even so, this remains an important factor to consider given the reliance of consolidation on memory, and should be examined in future work on the rehearsal of learning. There is also the alternate possibility that consolidation effects actually require longer breaks in training to be displayed: the breaks used in the present experiments may not have allowed for a sufficient number of replays to display any rehearsal benefit, even where replays were effective in altering the associative representation. This would be a more surprising finding given that the length of these breaks was based on those used in the experiments of Gershman et al. (2014), which did find behavioural differences across such relatively short intervals. It may then be advisable to replicate this study to determine whether such results are in fact reliable before applying this design to other tasks.

It is also worth noting that these two suggested aspects of Replay are likely to show significant interactions with one another: for example, the factors used to select replays could vary from task to task, while clearer definitions of replay selection criteria could help to better encourage rehearsal to counteract forgetting. These elements must therefore be considered simultaneously in order to provide a more complete depiction of learning consolidation. Both suggestions are still, however, merely interpretations of the common elements of these results rather than definitive findings, and will require further verification in future research using more varied tasks and designs.

4.5.1 Conclusion

The studies in this chapter aimed to examine the process by which an acquired representation could be re-examined and revalued in order to better capture external data patterns, focussing on the discovery of new concepts through the mental replay of past events. Instead, the results of these experiments show little evidence of a general benefit of rehearsal, with both contextual and temporal manipulations seemingly having little effect on behaviour. This may then depict a more complex consolidation system than initially offered by the current Replay model, involving deeper considerations of task demands and trial features to identify scenarios in which a benefit is observed. We therefore hope that further contrasts and applications to novel tasks will offer greater insight into this process, and a better understanding of the ways in which our representations are maintained.

Chapter 5

Conclusion

The preceding chapters of this thesis describe three studies each examining the form of our mental representations in varied tasks. While each of these tasks differ in subject and structure, the results of these studies do offer a consistent theme: the representations we build from our experiences are not simply a loose collection of event memories, but are crafted by internal systems to reflect external data patterns. This is demonstrated both in the clustering mechanisms used in the studies of stereotype use and numerosity of Chapters 2 and 3, as well as the connections formed between stimuli in the associative learning tasks of Chapter 4. Such findings suggest a general preference for structure in our learning systems, building advanced representations which reflect the complexity of our environment: categories are divided into distinct subgroups to reflect commonalities in members; numerical systems attempt to build distributions which reflect the prevalence of certain values in our experiences; and associative networks seek to reflect the costs and benefits of available actions to guide future decisions. What is more, these forms have clear consequences on resulting behaviour: exemplar partitioning determines subsequent stereotypical beliefs, numerical format affects discrete estimates, and associative structures direct related stimulus choices. These results then reveal the broader goals behind such systems, building representations which are not just accurate to true environmental structures, but also provide clear directions for related behaviour.

These studies also serve to demonstrate the value of computational models of behaviour in offering insight into such representations: all three studies presented here use such models both to provide descriptions of the process by which our representations are formed as well as to predict the outcomes of such systems on behaviour. This then allows for both qualitative and quantitative assessments of the mechanisms supporting our behaviour which may not possible using cognitive theory alone. It is important to note however that these models are not necessarily completely accurate; even where a model might receive support from comparisons with participant data, this is not definitive evidence of the use of this system by actual learners. A key factor in this distinction is model complexity: highly-complex models may provide good descriptions of behaviour, but may not be feasible within human capabilities. This reflects the position of these models within Marr's levels of description (Marr, 1982): the models used in these studies are predominantly computational, providing broad goals for behaviour rather than the actual implementation of these principles by real learners. These models will then likely require further development before they can be accepted as true depictions of human decision making, particularly regarding the algorithms required to make such structures feasible in actual learning. There are, however, multiple potential algorithms that could be applied to such models to fill this role according to the form of the representation; for example, sampling procedures can be used as an approximation for a number of Bayesian spatial methods such as the clustering models used here (Gelman et al., 2013), while network structures can use prediction errors to facilitate learning (Rumelhart, Hinton, & Williams, 1986). What is more, the use of such algorithms could also provide potential explanations for any systematic errors made by human learners in such tasks (Sanborn et al., 2010; M. Jones, Curran, Mozer, & Wilder, 2013), stepping away from the more complex systems described here.

It is also worth noting that the differences in representation discussed in each of these studies are focussed on the different potential outcomes of a single given method: Chapter 2 focuses on differences in exemplar clustering in stereotype change, Chapter 3 focuses on differences in prior format in numerical estimation,

and Chapter 4 focuses on differences in the strength of inter-stimulus associations in learning consolidation. While this does provide valuable distinctions in representational form in each of these studies, as was noted in the introduction to this thesis, there are a number of methods which have been applied as potential models of cognition with substantial differences in generated representation, ranging from logical rules to artificial neural networks. It may then be useful to extend the models considered in each of these domains to include such variations in methodology; for example, stereotypical beliefs could be examined using associative networks linking category membership to target traits, while learning trials could be partitioned using clustering techniques to infer latent structures, each providing alternate approaches to their given subject. Such applications could offer novel methods to examine the representations used in these tasks, including new descriptions of learning and new behavioural predictions to be tested in further work. One element to consider in such contrasts however is that these broader differences in representation between methods may not necessarily reflect actual differences in learning system, but could instead present different depictions of the same underlying process. This reflects the fact that these models provide different approaches for exploring the ways in which human behaviour operates, with no single model likely offering a perfect description of behaviour (Box, Hunter, & Hunter, 2005). Method selection may then depend not just on the match to human behaviour, but also the suitability of the representation used to the needs of the topic at hand; for example, clustering methods may be most useful when such categorisations are central to the target problem, as in Chapter 2, whereas the less transparent network methods may be better suited to tasks where the actions derived from a representation are more crucial than its form, as in Chapter 4.

Another aspect raised by such models comparisons is the optimality of behaviour: the Bayesian methods used in both the models of stereotyping and numerosity are commonly used to describe optimal solutions to a given problem (Anderson, 1991; Sanborn et al., 2010), while associative networks often attempt to identify the value of certain actions for the agent to determine optimal choices

(Sutton & Barto, 1981; Gershman et al., 2014). The optimality of human behaviour in contrast is far more questionable; many studies have noted the errors and fallacies displayed by human learners, thereby suggesting optimal processes may not accurately describe our actions (Tversky & Kahneman, 1974; Gigerenzer & Brighton, 2009). As noted above, however, such errors in human behaviour may be attributable to the specialised algorithms used to implement such high-level optimal systems, thereby introducing potential biases that generate the irrational behaviours observed in real life (Sanborn et al., 2010; M. Jones et al., 2013). This further reinforces the importance of considering such algorithms alongside higher-level goals when evaluating cognitive models, finding a combination which captures both the high and low levels of performance which can be demonstrated by real people.

One final point to make regarding the use of these models is on the potential advantages of the complexity assumed in such systems: while the above points question the feasibility of complex learning processes within human capabilities, it is also worth noting what benefits increased complexity may offer in capturing the intricacies of our behaviour. Recent advances in machine learning techniques have provided increasingly complex systems for use by artificial agents, and therefore cognitive models; these models can then capitalise on these advances to create more detailed behavioural descriptions, potentially providing better accounts for our own learning. This is particularly notable in modern deep learning systems, which have demonstrated great success in mirroring human levels of performance in real-world tasks such as image recognition (Farabet et al., 2013), speech processing (Hinton et al., 2012) and playing video and board games (Mnih et al., 2015; Silver et al., 2016). Continued application of such complex methods to models of human learning could then allow for more complete depictions of behaviour, shifting focus from the abstracted tasks commonly used in cognitive science to more valid problems matching with those encountered in everyday life. The use of such highly-complex models will however further require appropriate algorithms to make such systems feasible if they are to be applied as explanations for human learning.

To conclude, computational models of behaviour offer valuable insight into

the mechanisms we use to build and maintain our mental representations, providing potential explanations for both the strengths and weaknesses of human learning. This is here demonstrated in three distinct domains, each using such models to offer insight into our own learning processes, and the direct impact of the workings of these systems on our actions. The use of such models does however require a number of considerations before such systems can be accepted as accurate descriptions, particularly regarding the feasibility and implementation of what can be highly-complex methods. Continued development and comparison will therefore provide greater understanding of our own mental processes, and the ways in which we are able to make sense of our noisy and complex world.

List of Abbreviations

Chapter 2

BKM: Book-Keeping Model

SSM: Strong Subtyping Model

RRMC: Restricted Rational Model of Categorisation

RMC: Rational Model of Categorisation

BIC: Bayesian Information Criterion

Chapter 3

BDT: Bayesian Decision Theory

UEM: Uncertain Estimations Model

dUEM: Discrete Uncertain Estimations Model

cUEM: Continuous Uncertain Estimations Model

AIC: Akaike Information Criterion

Chapter 4

RW Model: Rescorla-Wagner Model

References

Acerbi, L., Vijayakumar, S., & Wolpert, D. M. (2014). On the origins of suboptimality in human probabilistic inference. *PLoS Computational Biology*, *10*, e1003661. doi: 10.1371/journal.pcbi.1003661

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. doi: 10.1109/TAC.1974.1100705

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII-1983* (p. 1-198). Berlin: Springer.

Allport, G. W. (1954). *The nature of prejudice*. Cambridge, Mass: Addison-Wesley.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. doi: 10.1037/0033-295X.98.3.409

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, *2*, 1152–1174. doi: 10.1214/aos/1176342871

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 89–195). New York: Academic Press.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*, 817–851. doi: 10.1037/a0034194

Bendor, D., & Wilson, M. A. (2012). Biasing the content of hippocampal replay during sleep. *Nature Neuroscience*, *15*, 1439–1444. doi: 10.1038/nn.3203

Biehl, M., Hammer, B., & Villman, T. (2016). Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews - Cognitive Science*, *7*, 92–111. doi: 10.1002/wcs.1378

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*, 7:1–7:30. doi: 10.1145/1667053.1667056

Born, J., Rasch, B., & Gais, H. (2006). Sleep to remember. *Neuroscientist*, *12*, 410–424. doi: 10.1177/1073858406292647

Bott, L., & Murphy, G. L. (2007). Subtyping as a knowledge preservation strategy in category learning. *Memory & Cognition*, *35*, 432–443. doi: 10.3758/BF03193283

Bowden, E. M. (1997). The effect of reportable and unreportable hints on anagram solution and the aha! experience. *Consciousness and Cognition*, *6*, 545–573. doi: 10.1006/ccog.1997.0325

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*, 322–328. doi: 10.1016/j.tics.2005.05.012

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (2005). *Statistics for experimenters (2nd ed.)*. Hoboken, NJ: Wiley-Interscience.

Brogden, W. J. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, *25*, 323–332. doi: 10.1037/h0058944

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576. doi: 10.1037/0033-295X.114.3.539

Bruner, S. J., Goodnow, J. J., & Austin, G. A. (1956). *The study of thinking*. New York, NY: Wiley.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action structure and causal relationships in continuous sequences of human action. *Cognitive Psychology*, *76*, 30–77. doi: 10.1016/j.cogpsych.2014.10.001

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A.,

... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Computational Biology*, *10*, e1003963. doi: 10.1371/journal.pcbi.1003963

Chalk, M., Seitz, A. R., & Series, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of Vision*, *10*, 2. doi: 10.1167/10.8.2

Chapman, G. B. (1991). Trial order affects cue interaction in contingency judgement. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *17*, 837–854. doi: 10.1037//0278-7393.17.5.837

Chen, D. P., & Mak, B. K. W. (2015). Multitask learning of deep neural networks for low-resource speech recognition. *IEEE-ACM Transactions on Audio Speech and Language Processing*, *23*, 1172–1183. doi: 10.1109/TASLP.2015.2422573

Cristianini, N., & Schölkopf, B. (2002). Support vector machines and kernel methods - the new generation of learning machines. *AI Magazine*, *23*, 31–41. doi: 10.1609/aimag.v23i3.1655

Davidson, T. J., Kloosterman, F., & Wilson, M. A. (2009). Hippocampal replay of extended experience. *Neuron*, *63*, 497–507. doi: 10.1016/j.neuron.2009.07.027

Decoste, D., & Schölkopf, B. (2002). Training invariant support vector machines. *Machine Learning*, *46*, 161–190. doi: 10.1023/A:1012454411458

Dehaene, S. (2011). *The number sense: how the mind creates mathematics*. New York: Oxford University Press.

Dehaene, S., & Marques, J. F. (2002). Cognitive euroscience: scalar variability in price estimation and the cognitive consequences of switching to the euro. *Quarterly Journal of Experimental Psychology*, *55*, 705-731. doi: 10.1080/02724980244000044

Ditz, H. M., & Nieder, A. (2016). Numerosity representations in crows obey the Weber-Fechner law. *Proceedings of the Royal Society B - Biological Sciences*, *283*, 20160083. doi: 10.1098/rspb.2016.0083

Erdogan, G., Yildirim, I., & Jacobs, R. A. (2015). From sensory signals to modality-

independent conceptual representations: A probabilistic language of thought approach. *PLOS Computational Biology*, *11*, e1004610. doi: 10.1371/journal.pcbi.1004610

Euston, D. R., Tatsuno, M., & McNaughton, B. L. (2007). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science*, *318*, 1147–1150. doi: 10.1126/science.1148979

Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 1915–1929. doi: 10.1109/TPAMI.2012.231

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*, 630–633. doi: 10.1038/35036586

Flombaum, J. I., Junge, J. A., & Hauser, M. D. (2005). Rhesus monkeys (Macaca mulatta) spontaneously compute addition operations over large numbers. *Cognition*, *97*, 315-325. doi: 10.1016/j.cognition.2004.09.004

Frank, M. C., & Tenenbaum, J. B. (2011). Three ideal observer models for rule learning in simple languages. *Cognition*, *120*, 360–371. doi: 10.1016/j.cognition.2010.10.005

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. London: CRC Press.

Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: a tale of two systems. *Journal of Experimental Psychology: General*, *143*, 182–194. doi: 10.1037/a0030844

Gershman, S. J., & Niv, Y. (2013). Perceptual estimation obeys occam's razor. *Frontiers in Psychology*, *4*, 623. doi: 10.3389/fpsyg.2013.00623

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*, 452–459. doi: 10.1038/nature14541

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143. doi: 10.1111/j.1756-8765.2008.01006.x

Gilhooly, K. J. (1978). Bigram statistics for 205 five-letter words having single-

solution anagrams. *Behavior Research Methods and Instrumentation*, *10*, 389–392. doi: 10.3758/BF03205158

Goldman, W. P., Wolters, N. C. W., & Winograd, E. (1992). A demonstration of incubation in anagram problem-solving. *Bulletin of the Psychonomic Society*, *30*, 36–38. doi: 10.3758/BF03330390

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A bayesian framework for word segmentation: exploring the effects of context. *Cognition*, *112*, 21–54. doi: 10.1016/j.cognition.2009.03.008

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154. doi: 10.1080/03640210701802071

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwinska, A., ... Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*, 471–476. doi: 10.1038/nature20101

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. (2007). Unifying rational models of categorization via the hierarchical dirichlet process. In R. Sun & N. Miyake (Eds.), *Proceedings of the 29th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Griffiths, T. L., & Ghahramani, Z. (2011). The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185–1224.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., Navarro, D. J., & Tenenbaum, J. B. (2011). Nonparametric Bayesian models of categorization. In E. M. Pothos & A. J. Wills (Eds.), *Formal approaches in categorization* (p. 173-198). Cambridge, UK: Cambridge University Press.

Griffiths, T. L., Vul, E., & Sanborn, A. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*, 263–268. doi: 10.1177/0963721412447619

Hayes, B. K., Banner, S., & Navarro, D. J. (2017). Sampling frames, Bayesian inference and inductive reasoning. In G. Gunzelmann, A. Howes, T. Tenbrink,

& E. J. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Heit, E., Briggs, J., & Bott, L. (2004). Modeling the effects of prior knowledge on learning incongruent features of category members. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *30*, 1065–1081. doi: 10.1037/0278-7393.30.5.1065

Heller, K., Sanborn, A. N., & Chater, N. (2009). Hierarchical learning of dimensional biases in human categorization. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22.* Cambridge, MA: MIT Press.

Hilton, J. L., & von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*, 237–271. doi: 10.1146/annurev.psych.47.1.237

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*, 82–97. doi: 10.1109/MSP.2012.2205597

Inoue, K., Ribeiro, T., & Sakama, C. (2014). Learning from interpretation transition. *Machine Learning*, *94*, 51–79. doi: 10.1007/s10994-013-5353-8

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*, 1221-1247. doi: 10.1016/j.cognition.2007.06.004

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, *51*, 343–358. doi: 10.1016/j.jmp.2007.06.002

Johnston, L., & Hewstone, M. (1992). Cognitive models of stereotype change: 3. subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology*, *28*, 360–386. doi: 10.1016/0022-1031(92)90051-K

Jones, J. L., Esber, G. R., McDannald, M. A., Gruber, A. J., Hernandez, A., Mirenzi, A., & Schoenbaum, G. (2012). Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science*, *338*, 953–956. doi:

10.1126/science.1227489

Jones, M., Curran, T., Mozer, M. C., & Wilder, M. H. (2013). Sequential effects in response time reveal learning mechanisms and event representations. *Psychological Review*, *120*, 628-666. doi: 10.1037/a0033180

Katzouris, N., Artikis, A., & Paliouras, G. (2015). Incremental learning of event definitions with inductive logic programming. *Machine Learning*, *100*, 555–585. doi: 10.1007/s10994-015-5512-1

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, *105*, 10687–10692. doi: 10.1073/pnas.0802631105

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, *10*, e1003915. doi: 10.1371/journal.pcbi.1003915

Kohonen, T. (2013). Essentials of the self-organizing map. *Neural Networks*, *37*, 52–65. doi: 10.1016/j.neunet.2012.09.018

Kool, W., & Botvinick, M. (2014). A labor/leisure trade-off in cognitive control. *Journal of Experimental Psychology: General*, *143*, 131–141. doi: 10.1037/a0031048

Kool, W., & Botvinick, M. (2018). Mental labour. *Nature Human Behaviour*, *2*, 899-908. doi: 10.1038/s41562-018-0401-9

Kording, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*, 244–247. doi: 10.1038/nature02169

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*, 84–90. doi: 10.1145/3065386

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgements. *Perception and Psychophysics*, *35*, 536–542. doi: 10.3758/BF03205949

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. doi: 10.1037/0033-

295X.99.1.22

Lake, B. M., Lawrence, N. D., & Tenenbaum, J. B. (2018). The emergence of organizing structure in conceptual representation. *Cognitive Science*, *42*, online pre-print. doi: 10.1111/cogs.12580

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*, 1332–1338. doi: 10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. doi: 10.1017/S0140525X16001837

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. doi: 10.1038/nature14539

Lubow, R. E. (1973). Latent inhibition. *Psychological Bulletin*, *79*, 398–407. doi: 10.1037/h0034425

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulleting and Review*, *22*, 1193–1215. doi: 10.3758/s13423-015-0808-5

Ludvig, E. A., Mirian, M. S., Kehoe, E. J., & Sutton, R. S. (2017). Associative learning from replayed experience. Online preprint. *biorxiv*, 100800.

Mansinghka, V., Shafto, P., Jonas, E., Petschulat, C., Gasner, M., & Tenenbaum, J. B. (2016). Crosscat: A fully bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *Journal of Machine Learning Research*, *17*, 138:1–49.

Maquet, P. (2001). The role of sleep in learning and memory. *Science*, *294*, 1048–1052. doi: 10.1126/science.1062856

Marr, D. (1982). *Vision: A computational investigation into the human representa-*

tion and processing of visual information. New York: Freeman.

Maurer, K. L., Park, B., & Rothbart, M. (1995). Subtyping versus subgrouping processes in stereotype representation. *Journal of Personality and Social Psychology*, *69*, 812–824. doi: 10.1037//0022-3514.69.5.812

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*, 348–356. doi: 10.1016/j.tics.2010.06.002

McCrink, K., & Wynn, K. (2004). Large-number addition and subtraction by 9-month-old infants. *Psychological Science*, *15*, 776-781. doi: 10.1111/j.0956-7976.2004.00755.x

McGaugh, J. L. (2000). Memory - a century of consolidation. *Science*, *287*, 248–251. doi: 10.1126/science.287.5451.248

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533. doi: 10.1038/nature14236

Momennejad, I., Otto, A. R., Daw, N. D., & Norman, K. A. (2018). Offline replay supports planning in human reinforcement learning. *Elife*, *7*, e32548. doi: 10.7554/eLife.32548

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519-1520. doi: 10.1038/2151519a0

Muggleton, S., Raedt, L. D., Poole, D., Bratko, I., Flach, P., Inoue, K., & Srinivasan, A. (2012). Ilp turns 20: Biography and future challenges. *Machine Learning*, *86*, 3–23. doi: 10.1007/s10994-011-5259-2

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, *27*, 148–193. doi: 10.1006/cogp.1994.1015

Nosofsky, R. M. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57. doi: 10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy,

S. M. Kosslyn, & R. M. Shiffrin (Eds.), *Essays in honor of william k. estes: Vol. 1. from learning theory to connectionist theory* (pp. 149–167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79. doi: 10.1037/0033-295X.101.1.53

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an exemplar-memory model of classification learning in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, *147*, 328–353. doi: 10.1037/xge0000369

Novick, L. R., & Sherman, S. J. (2003). On the nature of insight solutions: Evidence from skill differences in anagram solution. *Quarterly Journal of Experimental Psychology: Section A - Human Experimental Psychology*, *56*, 351–382. doi: 10.1080/02724980244000288

Penney, C. G., Godsell, A., Scott, A., & Balsom, R. (2004). Problem variables that promote incubation effects. *Journal of Creative Behavior*, *38*, 35–55. doi: 10.1002/j.2162-6057.2004.tb01230.x

Peterson, C. (1974). Incubation effects in anagram solution. *Bulletin of the Psychonomic Society*, *3*, 29–30. doi: 10.3758/BF03333382

Peterson, J., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th annual conference of the cognitive science society.* Austin, TX: Cognitive Science Society.

Pitman, J. (2002). *Combinatorial stochastic processes*. (Notes for Saint Flour Summer School).

Queller, S., & Mason, W. (2008). A decision bound categorization approach to the study of subtyping of atypical group members. *Social Cognition*, *26*, 66–101. doi: 10.1521/soco.2008.26.1.66

Rasmussen, M., Rieger, J., & Webster, K. N. (2017). Approximation of reachable sets using optimal control and support vector machines. *Jour-*

nal of Computational and Applied Mathematics, *311*, 68–83. doi: 10.1016/j.cam.2016.06.015

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308. doi: 10.1037//0033-295X.97.2.285

Razzaghi, T., Roderick, O., Safro, I., & Marko, N. (2016). Multilevel weighted support vector machine for classification on healthcare data with missing values. *PLOS One*, *11*, e0155119. doi: 10.1371/journal.pone.0155119

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407. doi: 10.1016/0010-0285(72)90014-X

Rescorla, R. A. (2004). Spontaneous recovery. *Learning and Memory*, *11*, 501–509. doi: 10.1101/lm.77504

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning ii.* New York: Appleton-Century-Crofts.

Richards, Z., & Hewstone, M. (2001). Subtyping and subgrouping: Processes for the prevention and promotion of stereotype change. *Personality and Social Psychology Review*, *5*, 52–73.

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and acquisition of language.* New York: Academic Press.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*, 386–408. doi: 10.1037/h0042519

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178–210. doi: 10.1006/jmps.2001.1379

Rouder, J. N., Speckman, P. L., Sun, D. C., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*, 225–237. doi: 10.3758/PBR.16.2.225

Rudoy, J. D., Voss, J. L., Westerberg, C. E., & Paller, K. A. (2009). Strengthening individual memories by reactivating them during sleep. *Science*, *326*, 1079–1079. doi: 10.1126/science.1179013

Rumelhart, D. E., Hinton, G., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533536. doi: 10.1038/323533a0

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing, volume 1: Foundations*. Cambridge, MA: MIT Press.

Sanborn, A. N., & Beierholm, U. R. (2016). Fast and accurate learning when making discrete numerical estimates. *PLoS Computational Biology*, *12*, e1004859. doi: 10.1371/journal.pcbi.1004859

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Science*, *20*, 883-893. doi: 10.1016/j.tics.2016.10.003

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167. doi: 10.1037/a0020511

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*, 411–437. doi: 10.1037/a0031912

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. doi: 10.1016/j.neunet.2014.09.003

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional biases in function learning. *Cognitive Psychology*, *99*, 44–79. doi: 10.1016/j.cogpsych.2017.11.002

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. doi: 10.1214/aos/1176344136

Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: teaching by, and learning from, examples. *Cognitive Psychology*, *71*, 55–89. doi: 10.1016/j.cogpsych.2013.12.004

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. doi: 10.1126/science.3629243

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1–42. doi: 10.1037/h0093825

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*, 484–489. doi: 10.1038/nature16961

Spelke, E. S., & Tsivkin, S. (2001). Language and number: a bilingual training study. *Cognition*, *78*, 45-88. doi: 10.1016/S0010-0277(00)00108-6

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature*, *437*, 1272–1278. doi: 10.1038/nature04286

Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, *88*, 135–170. doi: 10.1037/0033-295X.88.2.135

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*, 410–441. doi: 10.1037/rev0000052

Testolin, A., & Zorzi, M. (2016). Probabilistic models and generative neural networks: Towards a unified framework for modeling normal and impaired neurocognitive functions. *Frontiers in Computational Neuroscience*, *10*, 73. doi: 10.3389/fncom.2016.00073

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131. doi: 10.1126/science.185.4157.1124

Urcelay, G. P., Perelmuter, O., & Miller, R. R. (2008). Pavlovian backward conditioned inhibition in humans: Summation and retardation tests. *Behavioural Processes*, *77*, 299–305. doi: 10.1016/j.beproc.2007.07.003

Vanpaemel, W., & Storms, G. (2008). In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin and Review*, *15*, 732–749. doi: 10.3758/PBR.15.4.732

Vul, E., & Pashler, H. (2007). Incubation benefits only after people have been misdirected. *Memory and Cognition*, *35*, 701–710. doi: 10.3758/BF03193308

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*, *11*, 192–196. doi: 10.3758/BF03206482

Weber, R., & Crocker, J. (1983). Cognitive processes in the revision of stereotypic beliefs. *Journal of Personality and Social Psychology*, *45*, 961–977. doi: 10.1037/0022-3514.45.5.961

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*, 676–679. doi: 10.1126/science.8036517

Wimmer, G. E., & Shohamy, D. (2012). Preference by association: how memory mechanisms in the hippocampus bias decisions. *Science*, *338*, 270–273. doi: 10.1126/science.1223252

Appendix A

Additional Stereotype Change Results

This appendix provides additional testing of the empirical data described in Chapter 2, separating the previous analyses by trait congruency to examine any difference in results between stereotypical and counter-stereotypical trait ratings.

A.1 Separated Trait Types

Rather than aggregating ratings from congruent and incongruent traits into a single stereotypicality score, the following analyses instead examined the two trait types separately using matching tests. As with the main text, these analyses use Bayesian repeated measures ANOVAs for each experiment including the factors of test block and concentration condition, though ratings from the first test block were again excluded in all cases due to their role as a baseline unaffected by exposure to the respective exemplar set. Note that for incongruent traits, the predicted direction of effect for concentration is reversed, with ratings expected to be lower in the concentrated condition, as lower ratings indicate more stereotypical expectations.

A.1.1 Experiment 1

For congruent ratings from the first experiment, a significant effect was found for test block, $F(1,4) = 18.3$, $p < .001$, $BF_{inc} > 10000$, with ratings decreasing across the task, but no significant difference was found between concentration conditions, $F(1,4) = 1.22$, $p = .272$, $BF_{inc} = 0.31$, and no interaction was observed between these factors, $F(1,4) = 0.77$, $p = .545$, $BF_{inc} = 0.04$. Follow-up t-tests found a significant difference within the second test block, though Bayes factors suggest this did not meet the threshold for substantial evidence. These results are summarised in Table A.1.

Similar patterns were observed in incongruent ratings: while test block was again significant, $F(1,4) = 33.7$, $p < .001$, $BF_{inc} = \text{Inf}$, showing an increase in ratings over the experiment, no significant effect was found for concentration, $F(1,4) = 0.37$, $p = .544$, $BF_{inc} = 0.30$, and no significant interaction was observed between these factors, $F(1,4) = 1.92$, $p = .107$, $BF_{inc} = 0.20$. Follow-up t-tests meanwhile found a near-significant differences in ratings in the second and third test blocks, though neither met the level of substantial evidence, while all other test blocks indicated no difference in ratings between concentration conditions. These results are summarised in Table A.2.

A.1.2 Experiment 2

Within the second experiment, both trait types showed similar effects to those described in the main text. For congruent ratings, a significant effect was again found

Block	t	df	p	BF_{10}
1	0.19	114	0.849	0.20
2	1.80	114	0.037	1.62
3	1.07	114	0.143	0.56
4	1.06	114	0.146	0.55
5	0.40	114	0.346	0.27
6	0.51	114	0.305	0.30

Table A.1: Bayesian t-test results for congruent trait ratings from Experiment 1.

Block	t	df	p	BF_{10}
1	0.27	114	0.768	0.21
2	1.45	114	0.074	0.93
3	1.58	114	0.058	1.13
4	0.57	114	0.285	0.32
5	0.19	114	0.576	0.17
6	0.13	114	0.550	0.18

Table A.2: Bayesian t-test results for incongruent trait ratings from Experiment 1.

for test block, $F(1,4) = 13.8$, $p < .001$, $BF_{inc} > 10000$, with ratings decreasing across the task, but no significant effect was found for either concentration condition, $F(1,4) = 0.002$, $p = .968$, $BF_{inc} = 0.25$, or the interaction between concentration and test block, $F(1,4) = 1.87$, $p = .116$, $BF_{inc} = 0.21$. Subsequent t-tests again found no significant differences between conditions in any test block, with most blocks again showing substantial evidence of no difference, as shown in Table A.3.

Similarly, incongruent ratings from the second experiment showed a significant effect of test block, $F(1,4) = 24.5$, $p < .001$, $BF_{inc} > 10000$, increasing across exposure to the exemplar set, but no significant effects of concentration, $F(1,4) = 0.33$, $p = .569$, $BF_{inc} = 0.27$, or any interaction between these factors, $F(1,4) = 0.45$, $p = .772$, $BF_{inc} = 0.03$. Follow-up t-tests again found substantial evidence of no difference between conditions in all test blocks, as summarised in Table A.4

Block	t	df	p	BF_{10}
1	0.31	97	0.760	0.22
2	0.71	97	0.760	0.13
3	0.79	97	0.785	0.13
4	0.39	97	0.350	0.29
5	1.01	97	0.158	0.55
6	0.11	97	0.457	0.23

Table A.3: Bayesian t-test results for congruent trait ratings from Experiment 2.

Block	t	df	p	BF_{10}
1	1.03	97	0.305	0.33
2	0.94	97	0.824	0.12
3	0.83	97	0.795	0.13
4	0.23	97	0.592	0.18
5	0.59	97	0.721	0.14
6	0.12	97	0.549	0.19

Table A.4: Bayesian t-test results for incongruent trait ratings from Experiment 2.

A.1.3 Experiment 3

As with the previous experiment, separated trait types from Experiment 3 show similar results to the main analysis. Within congruent ratings, a significant effect was found for test block, $F(1,4) = 19.4$, $p < .001$, $BF_{inc} > 10000$, with ratings decreasing across the task, but no significant effect was found for concentration, $F(1,4) = 0.003$, $p = .957$, $BF_{inc} = 0.23$, or for the interaction between concentration and test block, $F(1,4) = 0.51$, $p = .730$, $BF_{inc} = 0.02$. Follow-up t-tests found substantial evidence for no difference between concentration conditions in any test block, shown in Table A.5.

Incongruent ratings show similar results, demonstrating a significant effect for test block, $F(1,4) = 25.0$, $p < .001$, $BF_{inc} > 10000$, with ratings increasing across the task, but no significant effect for concentration, $F(1,4) = 0.73$, $p = .393$, $BF_{inc} = 0.29$, or for the interaction between the two factors, $F(1,4) = 0.17$, $p = .954$, $BF_{inc} = 0.01$. As with congruent ratings, follow-up t-tests found substantial evidence for no difference in these ratings between concentration conditions in any test block, shown in Table A.6.

Block	t	df	p	BF_{10}
1	0.63	120	0.529	0.23
2	0.44	120	0.332	0.28
3	0.19	120	0.427	0.22
4	0.32	120	0.376	0.25
5	0.51	120	0.695	0.14
6	0.11	120	0.542	0.18

Table A.5: Bayesian t-test results for congruent trait ratings from Experiment 3.

Block	t	df	p	BF_{10}
1	0.47	120	0.643	0.21
2	0.72	120	0.765	0.12
3	0.88	120	0.809	0.11
4	0.71	120	0.760	0.18
5	0.56	120	0.712	0.13
6	1.02	120	0.845	0.10

Table A.6: Bayesian t-test results for incongruent trait ratings from Experiment 3.

Appendix B

Uncertain Estimation Model Results

This appendix gives additional details from the modelling exercise of Chapter 3, including alternative measures of model fit and further comparisons with lesioned versions of the model.

B.1 Additional Modelling Results

The following provides alternate model comparison results, beginning with the global fits assuming a common set of parameters across participants within each experiment, summarised in Table B.1. This does show a different finding to the individual fits in Experiment 1, with the dUEM having a better fit to behaviour in both measures, though as mentioned in the main text, fits are however substantially better when using individual parameters in both tasks, making those findings more helpful in separating the models.

Experiment	Model	MLL	AIC	*w*(AIC)	BIC	*w*(BIC)
Experiment 1	dUEM	-51506	103021	1	103052	1
	cUEM	-51661	103333	0	103381	0
Experiment 2	dUEM	-78718	157444	0	157477	0
	cUEM	-78485	156982	1	157032	1

Table B.1: Global modelling results from Experiments 1 and 2, where MLL is the maximum log likelihood for that model assuming common parameters across participants in each experiment.

Secondly, we here list the full results of the model comparisons using AIC values, summarised in Table B.2.

B.1.1 Experiment 1

As with the BIC scores above, aggregate AIC scores show the cUEM had a better fit to experimental data, though the number of participants best fit by each model was relatively even, being 17 for the dUEM and 22 for the cUEM. When divided by uncertainty condition, the cUEM better fit the 70% group, accounting for 14 of the 19 participants, while the dUEM better fit the 95% group, accounting for 12 of the 20 participants. In contrast with the BIC measures, this difference in ratio was confirmed to be significant, $\chi^2(1) = 4.50$, $p = .034$, suggesting behaviour did appear more continuous in the high-uncertainty condition.

B.1.2 Experiment 2

Aggregate AIC scores found the cUEM held a better fit to data from Experiment 2, accounting for 44 of the 60 participants. When divided by uncertainty condition, the cUEM held a better fit in both the 75% and 95% groups, suggesting behaviour was best described using a Gaussian mixture prior even where feedback is more reliable; this is further displayed in the ratios of participants best fit by each model, with the cUEM accounting for 23 of the 30 participants in the 70% condition and 21 of the 30 participants in the 95% condition. In contrast to the first experiment, this ratio did not significantly differ between groups, $\chi^2(1) = 0.34$, $p = .559$.

B.2 Model Lesioning

To test the actual impact of the use of these prior distributions on the accuracy of subsequent estimation, the continuous and discrete models described above were compared with a lesioned version of the UEM removing either prior, labelled the lUEM. This meant that responses were based solely on perceptual data, as defined

Experiment	Comparison	Model	MLL	AIC	w(AIC)
Experiment 1	Individual	dUEM	-43720	87752	0
		cUEM	-43461	87391	1
	70%	dUEM	-22335	44821	0
		cUEM	-22112	44453	1
	95%	dUEM	-21385	42930	0.982
		cUEM	-21349	42938	0.018
Experiment 2	Individual	dUEM	-66731	133941	0
		cUEM	-65994	132707	1
	70%	dUEM	-34740	69719	0
		cUEM	-34338	69037	1
	95%	dUEM	-31991	64222	0
		cUEM	-31655	63671	1

Table B.2: Alternate modelling results from Experiments 1 and 2, where MLL is the maximum log likelihood for that model, and w(AIC) is the weight of the AIC score for the given comparison between the discrete and continuous models.

by Equation 3.22, though this distribution was again modified by the response exponent and background distribution as in Equation 3.23:

$$p(R_t|X_t) = (1-w_b)\frac{logN(log(v_t),\sigma_l^2)^e}{\sum logN(log(v_t),\sigma_l^2)^e} + w_b U(v_1,v_2) \tag{B.1}$$

The dUEM, cUEM and lUEM were then run at the best fitting parameters found for each model for each participant in the above model comparison and used to calculate an estimate of accuracy by taking the average probability of the model giving the true displayed value as a response across estimate trials. The predicted accuracy of the lUEM was significantly lower than both the dUEM ($t(98) = 16.2$, $p < .001$) and cUEM ($t(98) = 14.8$, $p < .001$), suggesting the use of either the discrete or continuous prior distributions benefits estimation performance.