Categories, Variability, and Inferences: Essays on Human Judgment

Elizaveta Konovalova

TESI DOCTORAL UPF / ANY 2018

DIRECTOR DE LA TESI
Professor Gaël Le Mens Departament of Economics and Business

Acknowledgements

First of all, I would like to thank Gaël Le Mens for the support and, more importantly, guidance throughout my PhD. Gaël pushed me to be better, try harder and taught me to be the researcher that I am today. He inspired me to be curious, critical, and daring in my research. I am deeply grateful for the opportunity to work with him.

Second, I would like to express gratitude to many people who helped me to grow as a scientist, including Robin Hogarth, Mircea Epure, Gert Cornelissen, Christian Brownless, Jerker Denrell, Jose Apesteguia, Daniel Navarro Martinez. Your input was invaluable throughout my work on this thesis.

Third, the completion of this work would not have been possible without the support of my friends and family. I would like to thank Dmitry Khametshin, Stefan G. Gudmundsson, Federica Daniele, Paul Soto, Martí Guasch, Marlène Rump for willing to listen and help out during these 5 years. Also, thank you to Lily, Arkady, Alexander, and Natalia for being my family and encouraging me every step of the way. Special thank you to Irina. I know you would have been proud.

Finally, I would like to thank the Spanish Ministry of Economics and Competitiveness for providing funding for my studies (Grant PSI2013-41909-P to G. Le Mens). The following grants has funded the research reported in this thesis: Spanish Ministry of Economics and Competitiveness Grant #AEI/FEDER UE-PSI2016-75353, a Ramon y Cajal Fellowship (RYC-2014-15035), and a Grant IN[15]_EFG_ECO_2281 from the Fundación BBVA to G. Le Mens.

Abstract

The three chapters of this thesis explore how previous experience and mental categories shape human judgments. Chapter One provides a sampling explanation for the in-group heterogeneity effect - a tendency of people to perceive the groups they belong to as more heterogeneous than the groups to which they do not belong. It notes that because people are more likely to interact with the in-group members, they will experience more variability of the in-group than the out-group. Chapters 2 and 3 investigate how mental categories affect feature-based inferences when the category of the object, people perceive, is uncertain. In an influential paper, Anderson (1991) proposed a rational model for this task. The model proposes that the information from all the mental categories is integrated to make a prediction about unobserved features of the object. A crucial feature of this model is the conditional independence assumption – it assumes that the within-category feature correlation is zero. In prior research, this model has been found to provide a poor fit to participants' inferences. Chapter 2 argues that the failure of Anderson's rational model stems from the inconsistency between the task environment and the core assumption of the model. It notes that the studies reported in existing research relied on task environments without conditional independence and shows that when this assumption is satisfied, Anderson's model performs well. Chapter 3 proposes an extension of Anderson's model that allows mental categories to be characterized by feature correlations and shows that such general rational model provides a good fit to the existing inference data from experiments with uncertain categorization.

Resum

Els tres capítols de la tesi estudien com experiències prèvies i categories mentals poden determinar els judicis humans. El capítol 1 aporta una explicació a l'efecte heterogeni "in-group" - la tendència de la gent a percebre els grups als quals pertanyen com a més heterogenis que els grups als quals no pertanyen. L'estudi reporta que, com que la gent interactua més amb els membres del seu in-group, experimentaran més variabilitat en aquest in-group que en el seu out-grup. Els capítols 2 i 3 estudien com les categories mentals afecten les inferències basades en característiques quan la categoria de l'objecte, segons la gent, és incerta. En un article molt influent, Anderson (1991) va proposar un model racional per aquesta tasca. El model proposa que la informació de totes les categories mentals està integrada per fer una predicció sobre les característiques de l'objecte que no son observables. Una característica principal del model és el supòsit d'independència lineal - assumeix que dins de cada categoria, la correlació entre característiques és zero. En estudis previs, s'ha provat que aquest model sembla tenir un encaix pobre amb les inferències dels participants. El capítol 2 estableix que la problemàtica del model racional d'Anderson ve per la inconsistència entre el disseny de la tasca i el supòsit central del model. S'estableix que els estudis existents es basen en aquest disseny de tasca sense complir amb el supòsit d'independència lineal i es demostra que, quan aquest supòsit es compleix, el model d'Anderson funciona correctament. El capítol 3 proposa una extensió del model d'Anderson que permet caracteritzar les categories mentals amb característiques correlacionades entre sí i es prova que, aquest model racional encaixa correctament amb les dades provinent d'experiments existents en inferències sobre categories incertes.

Preface

In this thesis, I have developed two lines of research related to human judgment. More precisely, I was interested in different stages of the same process: how people learn about categories and use them. In the first project, I discussed the role of the environment in how people form judgments about social groups. In other words, I looked at how people form perceptions about different categories through sampling. The second project explores how people use these categories to make feature-inferences in a situation when the category is uncertain. Moreover, in both projects, I used the rational approach to cognition. The first project illustrates that even if people processed information correctly, their sampling behavior can lead to systematically biased perceptions about social groups. In the second project, I provide supportive evidence for the rational approach to feature-based inference with uncertain categorization. The thesis is divided into three chapters where the first chapter correspondents to the first line of research and chapters 2 and 3 correspond to the second.

The first chapter proposes an information sampling explanation for the in-group heterogeneity effect - a widely documented tendency of people to perceive the group they belong to as more heterogeneous than the group they don't belong to. I analyze a model in which an agent forms beliefs and attitudes about social groups from her experience. Consistent with robust evidence from social sciences, I assume that people are more likely to interact with in-group members than with out-group members. Therefore, they obtain larger samples of information about in-groups than about out-groups. Because estimators of variability tend to be right-skewed, but less so when the sample size is large, sampled in-group variability will tend to be higher than sampled out-group variability. This implies that even agents who process information correctly will be subject to the in-group heterogeneity effect. Using computer simulations, I demonstrate that this effect emerges under a wide range of assumptions about the structure of the environment and how experience translates into perceived group variability. The findings rely on the assumption that perceived group variability depends on sample variability. I provide evidence in support for this assumption by analyzing data from two nationally representative surveys, re-analyzing data from an existing experiment and a new experiment. My explanation suggests that the in-group heterogeneity effect is a consequence of the structure of the environment in which people live. It complements existing explanations that propose that information about in-group and out-group is processed differently.

The second and third chapters study how mental categories affect people's inferences. A key function of mental categories is to facilitate predictions about unobserved features of objects. At the same time, often people are uncertain about which category the object comes from. How to make inferences in such a situation has become the subject of a large body of research. Existing evidence, however, is mixed. Some studies suggest that when making inferences people use information only from the most probable category (Murphy & Ross, 1994), while others show that people ignore categories altogether and rely only on feature correlations

(O. Griffiths, Hayes, & Newell, 2012). These studies have an important limitation. They are based on an experimental paradigm with discrete-valued features. This implies that predictions of the model that ignores categories and one that makes optimal use of the categories are exactly the same.

Instead, I have designed a novel experimental paradigm that uses continuous features and allows for a clear distinction between different models. The second chapter uses this paradigm to provide supporting evidence for Anderson's rational model of feature inference which proposes that information from all the mental categories is integrated to make a prediction about unobserved features of the object (Anderson, 1991). Prior experiments concluded that Anderson's model does not explain well inferences under uncertain categorization. The chapter argues that this failure of the rational approach stems from the inconsistency between the task environment and the assumptions of Anderson's model. The model relies on an assumption that the within-category feature correlation is equal to zero. Yet none of the tasks in the existing studies are consistent with this assumption. Therefore, the aim of the paper was to measure the performance of Anderson's rational model in a setting where the within-category feature correlation is equal to zero. In five experiments, we found that Anderson's model outperforms competing models. This is an important finding because it suggests that when making category-based inferences, people are influenced by multiple categories.

In the third chapter, I propose an extension to Anderson's model to environments with the within-category correlations. I discuss the predictions of the rational model and other existing models in both discrete and continuous task environments. The analysis of existing findings and the results of a new experiment show that the rational model fits well the participants' inferences in both types of environments. Good performance of the rational model suggests that people use the category information optimally when making feature-based inferences with uncertain categorization.

Contents

List of Figures

List of Tables

Chapter 1

AN INFORMATION SAMPLING EXPLANATION FOR THE IN-GROUP HETEROGENEITY EFFECT

Joint with Gaël Le Mens

1.1 Introduction

A large amount of research has shown that people frequently perceive their groups as more heterogeneous than groups to which they do not belong (Boldry, Gaertner, & Quinn, 2007; Rubin & Badea, 2012; Ostrom & Sedikides, 1992). For example, Park and Judd (1990) found that students majoring in one subject judged students of other majors as less heterogenous on characteristics such as extroversion or impulsiveness. Linville, Fischer, and Salovey (1989) found that Yale undergraduate students perceived college students as more heterogeneous in friendliness than elderly people. By contrast, members of an elderly community in Florida perceived elderly people as more heterogeneous in friendliness than college students. This "in-group heterogeneity effect" has received several explanations. One type of explanation invokes differences in how information about in-groups and out-groups is processed (Ostrom & Sedikides, 1992; Ostrom, Carpenter, Sedikides, & Li, 1993; Park & Rothbart, 1982) or encoded (Judd & Park, 1988; Linville et al., 1989; Linville & Fischer, 1998; Park & Judd, 1990). The second type of explanation notes that people often have prior beliefs that the out-group is more homogenous than the in-group (Park & Hastie, 1987). The third type of explanation notes that the self is part of the in-group (Park & Judd, 1990). Because the self is frequently seen as distinctive, this would contribute to a perception of in-group heterogeneity. The fourth type of explanation takes as a premise that heterogeneity is seen as a

positive feature of social groups and that people want to have a more positive view of their in-groups than of out-groups. People would thus be motivated to see in-groups as more heterogeneous than out-groups (Ostrom & Sedikides, 1992; Rubin & Badea, 2012).

Here, we propose a distinct explanation for the in-group heterogeneity effect. We note that people tend to obtain larger samples of observations about in-groups than about out-groups. For example, people are more likely to interact with others of the same ethnicity, gender, social class, or occupation (Marsden, 1987; J. M. McPherson & Smith-Lovin, 1987; M. McPherson, Smith-Lovin, & Brashears, 2006). We show that this asymmetry in sample sizes implies the emergence of the in-group heterogeneity effect.

Key to our explanation is the observation that the variability of a sample of observations tends to increase with sample size. Consider for example the variance of a sample of k independent draws from a standard normal distribution (with mean $\mu = 0$ and variance $\sigma^2 = 1$). This *sample variance* is a random variable that can be written $\hat{\sigma}_k^2 = Q/(k-1)$ where Q is distributed according to a chi-squared distribution with $k-1$ degrees of freedom (χ^2_{k-1}). The mean of Q is $k-1$. Two features of chi-squared distribution are noteworthy: Q is right-skewed (the probability that the sample variance is lower than the mean is higher than 50%) and the skewness is decreasing in k (the skewness is equal to $\sqrt{8/(k-1)}$). Overall this implies that the sample variance tends to underestimate the true variance ($\sigma^2 = 1$): $P(\hat{\sigma}_k^2 < \sigma^2) > .5$. Crucially, the probability of underestimation *decreases with sample size*.

If people obtain larger samples about in-groups than about out-groups and the tendency to underestimate variability decreases with sample size, then the experienced variability of the in-group will tend to be larger than the experienced variability of the out-group. Under the assumption that the perceived heterogeneity of a group depends on the experienced variability of a group,[1] this implies that people will tend to perceive in-groups as more variable than out-groups.

This explanation for the in-group heterogeneity effect operates at a level different from the explanations mentioned at the beginning of the introduction. Whereas these focus on how the mind processes information, our explanation emphasizes the properties of the information samples on which the mind operates – the input of mental operations (Brunswick, 1952; Fiedler & Juslin, 2006a; H. A. Simon, 1956).

Previous research has noted the importance of sample size in estimations of variability, in general (Kareev, Arnon, & Horwitz-Zeliger, 2002), and in the context of the in-group heterogeneity effect (Linville et al., 1989). But the theoretical arguments developed in these papers differ from ours. They focused on the properties of uncorrected sample variance as a statistical estimator of the variance of a distribution (σ^2):

$$\hat{\sigma}_k^2 = \frac{1}{k}\sum_{j=1}^{k}(x_j - \bar{x}_k)^2, \tag{1.1}$$

[1] We review and provide new evidence supporting this assumption in a latter section.

where $\bar{x}_k$ is the sample mean. They noted that this estimator is statistically negatively biased, especially when based on a small sample. To see this, consider the formula for the mean of the uncorrected sample variance:

$$E[\hat{\sigma}_k^2] = \sigma^2 - \frac{1}{k}\sigma^2 < \sigma^2. \tag{1.2}$$

If people's perception of group heterogeneity corresponds to the uncorrected sample variance, they will tend to underestimate the true variability. And the amplitude of the underestimation will diminish as sample size increases. If people obtain larger samples about the in-group than about the out-group, this implies: $E[\hat{\sigma}_{in}^2] > E[\hat{\sigma}_{out}^2]$, where $\hat{\sigma}_{in}^2$ and $\hat{\sigma}_{out}^2$ denote the sample variances of the two groups.

Even though this argument is elegant and parsimonious, its scope is limited by the fact that it only works for biased estimators of variability such as the uncorrected sample variance or the probability of differentiation (also analyzed by Linville et al., 1989). Moreover, according to currently available empirical evidence, it is unclear to what extent intuitive perceptions of variability correspond to these estimators.[2]

We propose a new perspective on how an asymmetry in sample sizes can explain the in-group heterogeneity effect. Our sampling argument works for many estimators of variability, including those considered by Linville et al. (1989). By contrast to these authors, our focus is not on the means of the sample variance distributions nor on the comparison of these means. Instead, our focus is on the probability that the in-group will be perceived as more variable than the out-group. This seemingly minor change has far-ranging consequences: it considerably expands the scope and relevance of sampling explanations for the in-group heterogeneity effect.

Consider an agent and her perceived variability for the in-group, V_{in}, and the out-group, V_{out}. We are interested in $P(V_{in} > V_{out})$. There is an in-group heterogeneity effect whenever the probability of perceiving the in-group as *more* variable than the out-group is *larger* than the probability of perceiving it as *less* variable than the in-group:

$$P(V_{in} > V_{out}) > P(V_{in} < V_{out}).$$

From a theoretical standpoint, the main contribution of this paper is to show that the structure of the environment is sufficient to explain the emergence of the in-group heterogeneity effect: even perceivers who would process information correctly (even rationally) would tend to perceive the in-group as more variable than the out-group. We do not claim that people process information rationally. Rather, we rely on the assumption of rational information processing to isolate the role

[2] Although prior literature has noted that most estimators of variability are highly correlated (Pollard (1984), cited in Kareev et al., 2002), the argument that focuses on the statistically biased nature of the estimator does not apply to unbiased estimators, such as the corrected sample variance. Corrected sample variance and uncorrected sample variance are highly correlated estimators, yet, the argument only works for the *uncorrected* sample variance.

of the environment, following the precepts of the rational analysis of cognition described by Anderson (1991). Like other rational analyses, our approach emphasizes how the structure of the environment can lead to systematic information asymmetries, which in turn imply systematic judgment asymmetries (Brunswick, 1952; Gigerenzer, Todd, & Group, 1999; Gigerenzer & Selten, 2002; Hogarth & Karelaia, 2007; Le Mens & Denrell, 2011; H. A. Simon, 1956). And because it focuses on properties of the information samples to which people have access, our explanation contributes to the 'sampling approach' to human judgment (Denrell, 2005; Einhorn & Hogarth, 1978; Fazio, Eiser, & Shook, 2004; Fiedler, 2000; Fiedler & Juslin, 2006a; Galesic, Olsson, & Rieskamp, 2012; Kareev, 2000; Le Mens & Denrell, 2011; Le Mens, Kareev, & Avrahami, 2016; March, 1996).

In what follows, we first describe our model and report computer simulations that show how the in-group heterogeneity effect emerges when perceived variability is assumed to be the (unbiased) corrected sample variance estimator (Section 'Simple Model'). We show that a similar pattern emerges when the measure of group variability is a Bayesian estimator of variance. In the section 'Other Measures of Variability', we demonstrate that a similar pattern emerges for a number of other variability estimators used in the empirical literature on the in-group heterogeneity effect. Following the presentation of these results, we discuss how our findings relate to other explanations for the in-group heterogeneity effect (Section 'Relation to Prior Explanations of In-Group Heterogeneity Effect'). Then, we note that several papers have documented a seemingly opposite empirical pattern: an in-group homogeneity effect (Section 'In-group Homogeneity Effect'). We show how our sampling-based mechanism can reconcile the findings about the in-group homogeneity effect and the findings about the out-group homogeneity effect. The gist of our explanation is that the two sets of findings concern different types of environments. The different environment structures imply systematic differences in the nature of the relevant information samples available to people. This, in turn, implies a systematic difference in the propensity to perceive the in-group as more variable or less variable than the out-group. In the following three sections, we discuss what happens under alternative assumptions about the sampling mechanisms. First, we note that our basic model implicitly assumed that groups were of infinite sizes. We analyze a model with finite group sizes (Section 'The Role of Group Size'). Next, we relax the assumption that people have perfect memory for the observations they collected. We analyze a model where people form variability estimates by sampling observations from memory (Section 'Sampling from Memory'). In the section 'Hedonic Sampling', we relax our model assumption that the sampling probabilities of the groups are fixed. We study what happens when sampling is motivated by a hedonic goal (e.g., Thorndike, 1927; see Denrell, 2005 for a review). In the section 'Variability of Perceived and Sampled Distributions,' we review the existing empirical evidence supporting our assumption that perceived group variability depends on sample variability. We discuss existing experimental findings, analyze data from two nationally representative surveys, re-analyze data from an existing experiment. We also report on a new experiment designed to

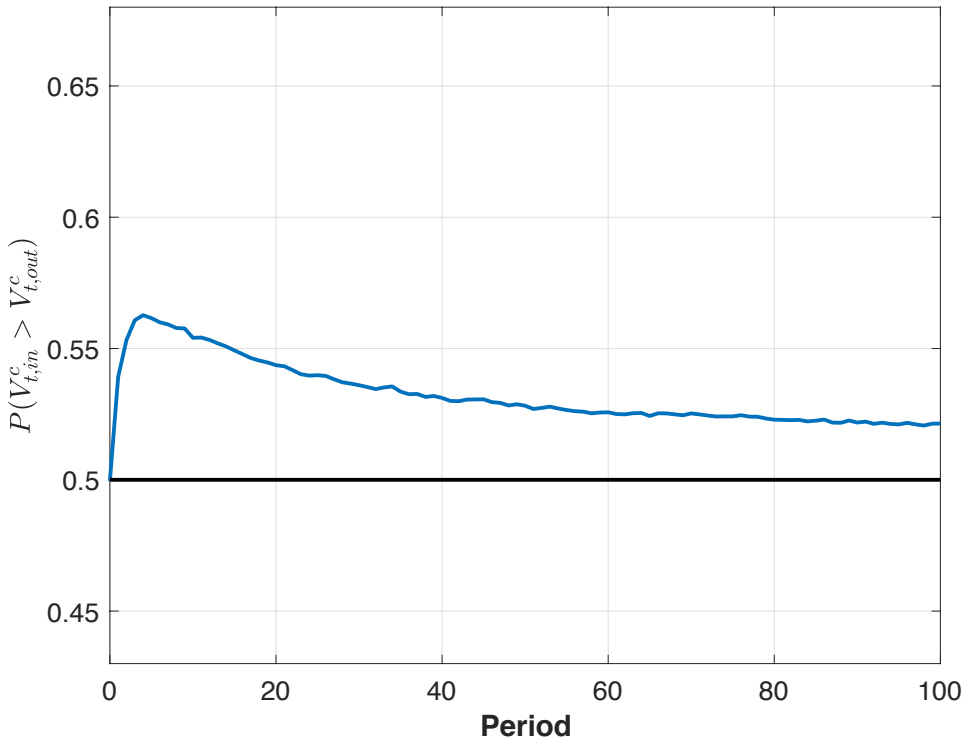


Figure 1.1: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(V^c_{t,in} > V^c_{t,out})$, when the estimator of variability is the corrected sample variance (eq. 1.3). Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma^2_{in} = \sigma^2_{out} = 1$.

complement the evidence found in earlier literatures. Finally, in the section 'Theoretical Implications', we note that our analyses can be seen as a 'rational analysis of cognition' (Anderson, 1991). We discuss what it implies for the possibility of correcting the in-group heterogeneity effect and how our findings relate to prior rational analyses.

1.2 Simple Model

We analyze a model in which an agent forms variability estimates of two groups on the basis of the samples she collects from the groups. We designed the model to be the simplest model that would illustrate the emergence of the in-group heterogeneity effect as a result of a difference in the sizes of the samples collected about the two groups. In later sections, we revisit our model assumptions and show that the basic result holds under a wide set of specifications.

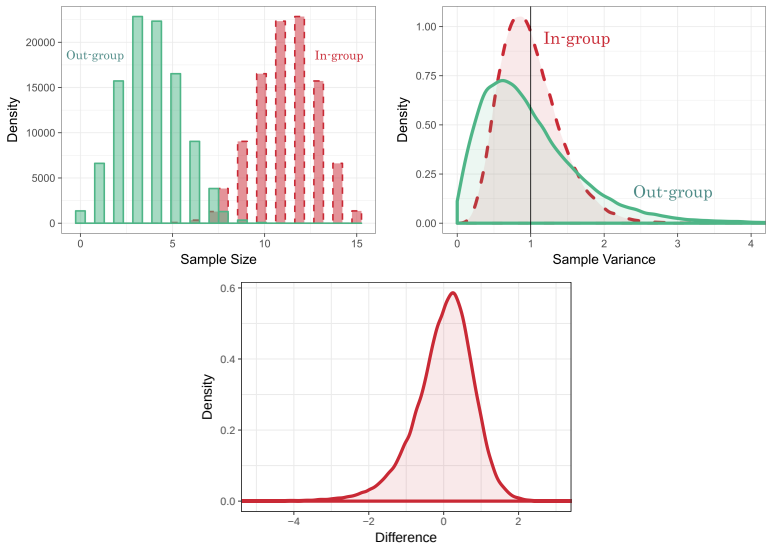


Figure 1.2: Model with corrected sample variance at the end of period 15. *Left Panel:* Distribution of the sample sizes of the two groups. *Middle Panel:* Distribution of variability estimates for the two groups, $V^c_{t,in}$ and $V^c_{t,out}$. The black vertical line denotes the true variance. *Right Panel:* Distribution of difference in variability estimates $\Delta V^c_t = V^c_{t,in} - V^c_{t,out}$. Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma^2_{in} = \sigma^2_{out} = 1$.

1.2.1 Model Description

Consider a setting where one agent forms beliefs about two groups ($g = in, out$). The agent belongs to one of the two groups – the in-group. In this simple model, we assume the agent observes just one dimension of the groups. We call it X. The two groups have the same variability on this dimension. In each period, the agent samples the group or not. When the agent samples a group she observes the X value of one of its members.

Distribution of the Observed Values The focal dimension has the same distribution in the two groups. In the baseline model, we assume this distribution to be Normal, with mean 0 and variance 1: $\mu_{in} = \mu_{out} = 0$, $\sigma^2_{in} = \sigma^2_{out} = 1$.

Sampling Rule To ensure that variability estimates exist for both groups, we assume that the agent has sampled 2 observations from each group before the first period (to keep the formulas as simple as possible, we assume they are done in periods -1 and 0). In each period $t \geq 1$, the agent samples the in-group or the out-group. Let r characterize the probability that the agent samples the in-group. We assume that r is larger than 0.5: the agent is more likely to sample the in-group than the out-group. If she samples group g, she obtains an observation $x_{t,g}$.

If she does not sample this group she does not obtain any additional observation. This sampling advantage for the in-group implies that the agent will gather larger samples of information about the in-group than about the out-group.

Perception of Variability We assume that the agent processes sampled information correctly: the agent has perfect memory of all the observed samples and uses a statistically unbiased estimator of variability. Let $V_{t,g}$ denote the perceived variability on dimension X at the end of period t. We assume that this is given by the corrected (unbiased) sample variance:

$$V^c_{t,g} = \frac{1}{n_{t,g}-1} \sum_{j=-1}^{t} (x_{j,g} - \bar{x}_{t,g})^2 I_{j,g}, \tag{1.3}$$

where $I_{j,g}$ is an indicator variable equal to 1 if group g is sampled in period j and equal to 0 otherwise, $n_{t,g}$ is the number of samples ($n_{t,g} = 2 + \sum_{j=1}^{t} I_{j,g}$), $\bar{x}_{t,g}$ is the mean of the sampled observations at the end of period t, and $x_{j,g}$ is the observation in period j.

Summary The agent in our model is subject to an environmental constraint that makes her sample the in-group more frequently than the out-group. She is a rational processor of information in that she has perfect memory and uses a minimum-variance unbiased estimator of variability.[3]

1.2.2 Analysis

Baseline Setting

We ran computer simulations of the model with $r = .75$ (the agent is three times more likely to sample the in-group than the out-group). Figure 1.1 displays the likelihood that the estimate of the in-group variability is higher than the estimate of the out-group variability $P(V^c_{t,in} > V^c_{t,out})$ as a function of the number of periods. It is higher than 0.5 for all periods after period 1. In other words, the in-group tends to be perceived as more variable than the out-group even though the true variabilities are the same.

The likelihood first increases quickly and then decreases slowly with the number of periods. This asymmetry persists for a large number of periods. Even after 50 or even 100 periods the probability is still higher than 0.5 (it is 0.52 after 100 periods).

[3]The corrected sample variance is an unbiased estimator of variance that minimizes mean square error among unbiased estimator when the underlying distribution is a Normal distribution. In that sense, it is the 'best' estimator (Casella & Berger, 2002).

1.2.3 Intuition

To develop an intuition for this result, it is useful to examine what happens at a specific point in time. We focus on the end of period 15. Specifically, we analyze the distributions of the corrected sample variances for the two groups. First note that the in-group is sampled more times than the out-group (Figure 1.2, *Left Panel*). This is because of the assumed sampling advantage of the in-group ($r = .75$). Second, note that the distributions of sampled variabilities for the two groups are right skewed but to a *different extent* (Figure 1.2, *Middle Panel*). The distribution of the sample variance of the in-group $V^c_{15,in}$ is less skewed than the distribution of the sample variance of the out-group $V^c_{15,out}$. Overall, this implies that $V^c_{15,in}$ tends to be larger than $V^c_{15,out}$, as shown by the distribution of $\Delta V^c_t = V^c_{t,in} - V^c_{t,out}$ (Figure 1.2, *Right Panel*). We have $P(V^c_{15,in} > V^c_{15,out}) = .55$. It is worth noting that the mean sample variances are the same and equal to the true variance: $E(V^c_{15,in}) = E(V^c_{15,out}) = 1$. This is because the corrected sample variance is (by design) a statistically unbiased estimator. Yet, an in-group heterogeneity effect emerges: most simulated agents experience the in-group as more variable than the out-group.

1.2.4 Sensitivity to Model Parameters

Baseline Probability of Sampling from the In-Group

The probability of sampling from the in-group (parameter r) reflects the tendency to interact more frequently with others of the same social group than with others of different social groups. Figure 1.3 illustrates the likelihood $P(V^c_{t,in} > V^c_{t,out})$ as a function of r. For all values of $r > 0.5$, an in-group heterogeneity effect emerges: $P(V^c_{t,in} > V^c_{t,out}) > .5$. Not surprisingly, the opposite effect emerges when $r < 0.5$. In this case, and in-group *homogeneity* effect emerges: $P(V^c_{t,in} > V^c_{t,out}) < .5$. We return to the possibility of an in-group homogeneity effect in a later section.

The extent to which people sample the in-group more often than the out-group (the value of r) depends on aspects of the social environment such as racial or ethnic segregation or the degree of homophily in people's social networks. Currently available evidence suggests that in many environments, people predominantly interact with others of the same group (see Denrell, 2005, for a review). For example, in many cities in the US and elsewhere, there is spatial segregation based on ethnicity (Van Kempen & Şule Özüekren, 1998) or race (Massey & Denton, 1989): the immediate social environment of people generally consists of others of the same race or ethnicity. Analyses of social networks based on data collected in nationally representative panels of respondents indicate that this tendency is widespread (Marsden, 1987; M. McPherson, Smith-Lovin, & Cook, 2001; M. McPherson et al., 2006). In terms of our model, this suggests that many environments correspond to $r > .5$. Yet, in some environments, people might be more likely to sample the out-group than the in-group ($r < .5$). This is likely to be the case for people who are

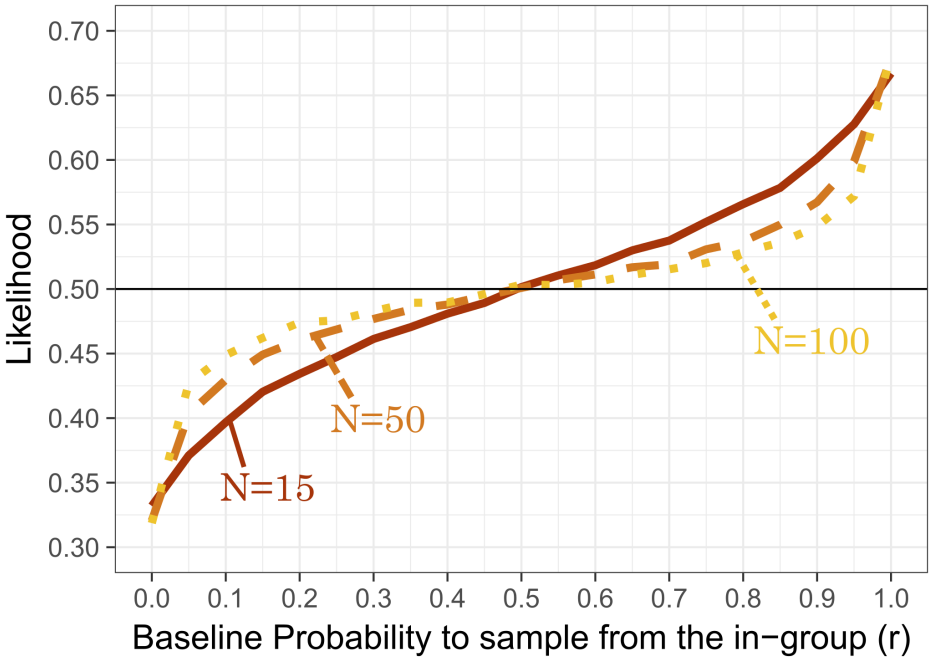


Figure 1.3: Likelihoods that the estimate of in-group variability is higher than the estimate of out-group variability, $P(V^c_{t,in} > V^c_{t,out})$, after 15 (solid line), 50 (dashed line) and 100 periods (dotted line), as a function of the baseline probability of sampling the in-group (r). The estimator of variability is the corrected sample variance (eq. 1.3). Based on 10^5 simulations with $\mu_{in} = \mu_{out} = 0; \sigma^2_{in} = \sigma^2_{out} = 1$.

part of a minority. Our model predicts that in this case, an in-group homogeneity effect will emerge.

In the analysis of the baseline model, we used $r = 0.75$. This means that the individual is three times more likely to sample from the in-group than from the out-group. Although it is difficult to obtain reliable data about the frequency of inter-group interactions, this number is consistent with empirical data. We analyzed the racial composition of communities in the United States using the 2000 edition of General Social Survey. These data are collected by the National Opinion Research Center at the University of Chicago and is based on a representative sample of US citizens (Davis, Smith, & Marsden, 2016). In the survey, respondents were asked to report their race and indicate the percentage of different races and ethnicities in their communities. Because of the specific format of the racial identity question, we estimated the community structure only for white and black respondents.[4] We calculated the share of the reported percentage of the in-group

[4]Although respondents were asked about other races and ethnicities in their communities, we were unable to assess the community structure for them. This is because when asked about their own race, the respondents were given three choices 'White', 'Black', and 'Other'. This formulation

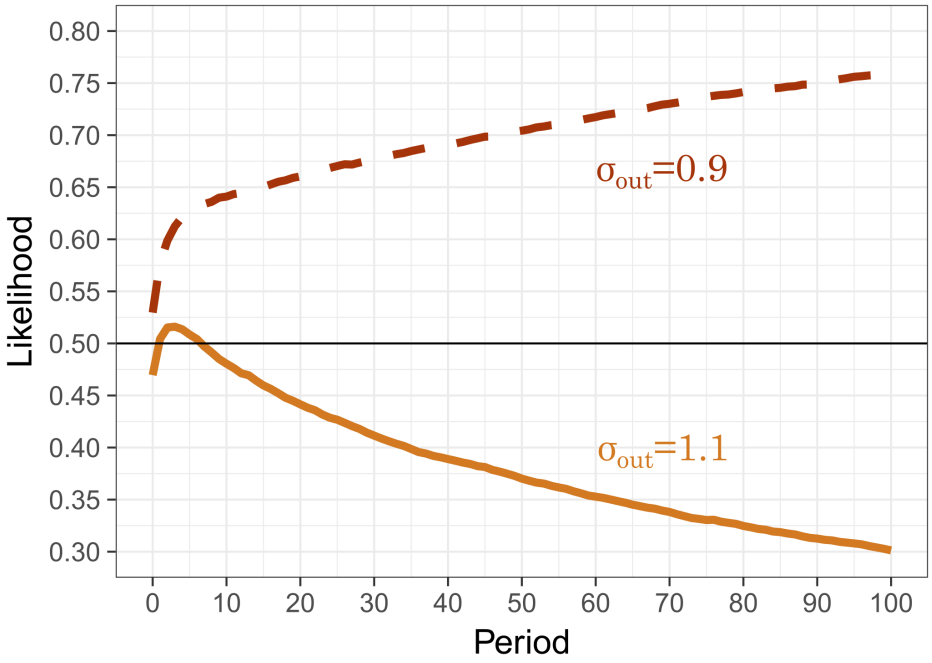


Figure 1.4: Likelihoods that the estimate of in-group variability is higher than the estimate of out-group variability, $P(V^c_{t,in} > V^c_{t,out})$, for two levels of the true variability of the out-group: $\sigma_{out} = 0.9$, $\sigma_{out} = 1.1$. For both cases, the true variability of the in-group is $\sigma_{in} = 1$. The estimator of variability is the corrected sample variance (eq. 1.3). Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0$.

members in the combined reported percentages of in-group and out-group members. For example, if a white person reported that his or her community is 50% white and 20% black, the resulting estimate is $\frac{50}{50+20} = 0.71$. The average estimates across all black and white respondents are 0.82 and 0.54 respectively. The median values are a bit higher at 0.88 and 0.60 respectively. These estimates indicate that white people have more than 4 times more whites than blacks in their communities. The asymmetry is not as large for black people. This is not surprising because the group of white people is the majority group in the US whereas the group of black people is a minority group. We explore issues related to group sizes in a latter section.

Difference in True Variabilities

To illustrate the implications of our sampling-based mechanism, we assumed that the true variabilities of the two groups were the same ($\sigma_{in} = \sigma_{out}$). This does not

makes the identification of the in-group other than 'White' and 'Black' impossible.

have to be the case. If the variabilities are different, our mechanism will imply the emergence of errors in the perceived difference in variability. Suppose the variability of the in-group is higher than the variability of the out-group: $\sigma^2_{in} > \sigma^2_{out}$. In this case, most agents will perceive the difference in variabilities as higher than what it really is: $P(V^c_{t,in} - V^c_{t,out} > \sigma^2_{in} - \sigma^2_{out}) > .5$. Moreover, the proportion of agents who perceive the in-group as more variable than the out-group will be higher than when the true variabilities are the same. As an illustration, suppose $\sigma_{in} = 1 > \sigma_{out} = 0.9$ (see Figure 1.4). After 15 periods, we have $P(V^c_{15,in} - V^c_{15,out} > \sigma^2_{in} - \sigma^2_{out}) = .53$. The probability of perceiving the in-group as more variable is $P(V^c_{15,in} - V^c_{15,out} > 0) = .65$. It was .55 when the true variabilities were the same.

Suppose now that the variability of the in-group is lower than the variability of the out-group: $\sigma_{in} < \sigma_{out}$. Just as before, most agents will perceive the difference in variabilities as higher than what it is: $P(V^c_{t,in} - V^c_{t,out} > \sigma^2_{in} - \sigma^2_{out}) > .5$. But an in-group *homogeneity* effect (rather than an in-group *heterogeneity* effect) could emerge if the difference in true variabilities is large enough: $P(V^c_{t,in} - V^c_{t,out} > 0) < .5$. Suppose $\sigma_{in} = 1 < \sigma_{out} = 1.1$ (see Figure 1.4). After 15 periods, we have $P(V^c_{15,in} - V^c_{15,out} > \sigma^2_{in} - \sigma^2_{out}) = .56$. This probability is similar to what we had before. But an in-group *homogeneity* effect emerges: $P(V^c_{t,in} - V^c_{t,out} > 0) = .46$. We return to the difference in true variabilities in the section on the in-group *homogeneity* effect.

1.2.5 Distribution of the focal feature

The nature of the asymmetry in perceived group variabilities depends on the nature of the distribution of the focal feature. We assumed it was normally distributed. Ancillary simulations suggest that the in-group heterogeneity effect emerges with most unimodal distributions (although we could not prove formally that unimodality is a sufficient condition for the effect to emerge). When the distribution is not unimodal, an opposite effect can emerge: an in-group *homogeneity* effect. The intuition for this result is that when the distribution is bimodal, sample variance does not increase with sample size, but instead tends to decrease. Suppose the distribution of the focal feature is a $Beta(0.2, 0.2)$ (for both groups). After 15 periods, we have $P(V^c_{15,in} > V^c_{15,out}) = .47$ (with $r = .75$). What is noteworthy is that this quantity is lower than .5. The support of the $Beta(0.2, 0.2)$ distribution is $[0, 1]$. It has a peak at 0 and a peak at 1. When the sample of observations is small, it is likely that all observations will be close to one of the two peaks. But as the sample size increases, it is more likely that some intermediary values (close to the mid-point, 0.5) will be sampled. The sample variance thus decreases. We are not aware of any existing study of the in-group heterogeneity effect that focused on features with bimodal distributions. But this distinctive prediction of our model could potentially be empirically tested.

1.2.6 Bayesian Estimator of Variance

We assumed that the perceived variability was the (unbiased) corrected sample variance. We used this estimator because we wanted to show that an in-group heterogeneity effect can emerge even if the information is processed 'correctly'. Another way to implement the idea of 'correct processing of information' in our model is to assume the agent is Bayesian and knows the structure of the environment (i.e., possesses correct priors).

Suppose the true variances are drawn from a distribution known to the agent. This distribution is her *prior* on the variances. The agent updates this prior based on her observations of the two groups using Bayes' theorem. For simplicity, we assume that the mean on the X dimension is *known* and equal to 0.[5] The true variances for the two groups σ^2_{in} and σ^2_{out} are both drawn from a uniform distribution $U(0,1)$. With this assumption the probability that the in-group is truly more variable than the out-group is .5: $P(\sigma^2_{in} > \sigma^2_{out}) = P(\sigma^2_{in} < \sigma^2_{out}) = .5$.

We denote by $V^{Bayes}_{t,g}$ the Bayesian estimator of the variance for a group g at the end of period t. This is the mean of the posterior distribution of the variance of the focal feature in group g. We computed the mean of the posterior using the Markov Chain Monte Carlo method explained in Shi, Griffiths, Feldman, and Sanborn (2010).

We simulated our model by substituting the corrected sample variance estimator by the Bayesian estimator. Figure 1.5 displays the likelihood that the estimate of the in-group variability is higher than the estimate of the out-group variability $P(V^{Bayes}_{t,in} > V^{Bayes}_{t,out})$ as a function of the number of periods. It is higher than 0.5 for all periods after period 1. In other words, the in-group tends to be perceived as more variable than the out-group. The effect persists even after a large number of periods.

It is important to note that the Bayesian estimator of variance is unbiased. To see this, note that the means of the distributions of the true variances are

$$E(\sigma^2_{in}) = E(\sigma^2_{out}) = 0.5.$$

After any period t, the means of the posteriors remain the same:

$$E(V^{Bayes}_{t,in}) = E(V^{Bayes}_{t,out}) = 0.5.$$

Our analyses considered an environment where the two groups were equally likely to be the more variable (the two variances were drawn from the same distribution). We showed that even when the samples are processed using Bayesian updating and the agent possesses correct priors, most simulated agents will perceive the in-group as more variable than the out-group – the in-group heterogeneity effect emerges.

[5] Similar results hold if the means are unknown.

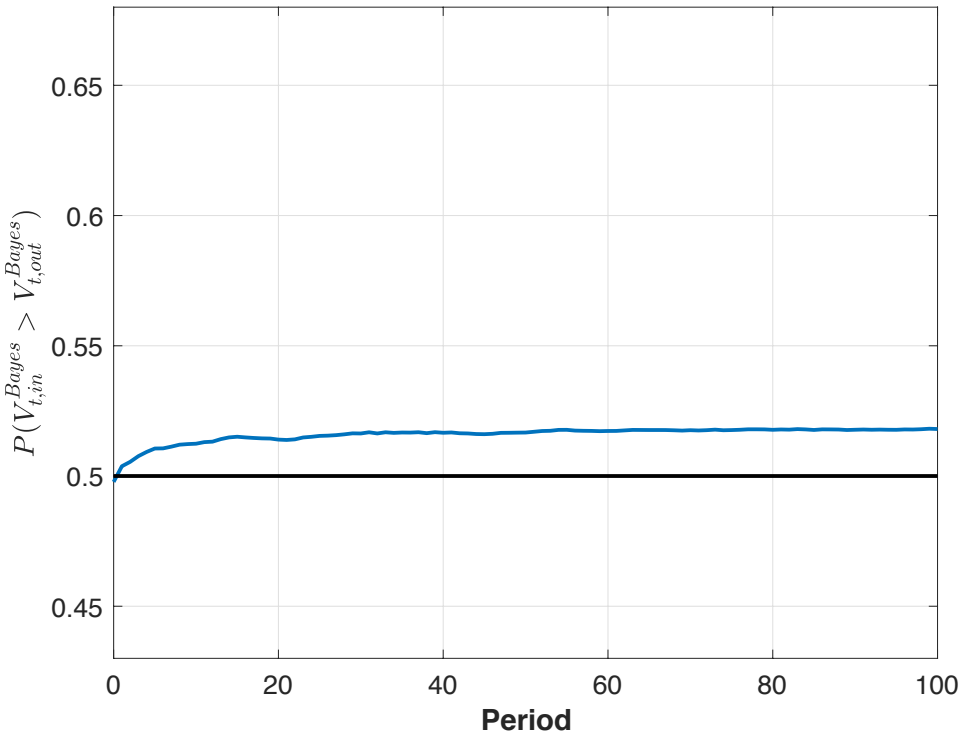


Figure 1.5: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(V_{t,in}^{Bayes} > V_{t,out}^{Bayes})$, when the measure of variability is the Bayesian Posterior. Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma_{in} \sim U(0,1), \sigma_{out} \sim U(0,1)$.

1.2.7 Discussion

When agents obtain larger samples of observations about the in-group than about the out-group, most agents will experience the in-group as more variable than the out-group even if it is not more variable. This asymmetry in experienced variability implies that most agents will perceive the in-group as more variable than the out-group, even if they use unbiased estimators of variance such as the corrected sample variance or a Bayesian estimator of variance. These findings are important from a theoretical standpoint because they demonstrate that the in-group heterogeneity effect can be explained by the structure of the environment rather than by invoking characteristics of how information is processed. We return to this issue in the 'Theoretical Implications' section.

Next, we demonstrate that similar results hold under alternative assumptions – possibly with greater psychological realism – about how experience translates into perceived variability. We analyze the emergence of the in-group heterogeneity effect when perceived variability is measured by other estimators used in the prior literature on the in-group heterogeneity effect.

1.3 Other Measures of Variability

Even though sample variance is the measure of variability that possibly most naturally comes to mind (of researchers), the extent to which it is a good estimator of *intuitive* perception of variability is unclear (Hogarth, 1975). Early studies note that central tendency and extreme observations are more salient to subjects (Hamilos & Pitz, 1977) and that subjects weight smaller deviations more than larger ones (Beach & Scopp, 1968). Furthermore, Peterson and Beach (1967) showed that variability judgments increase with sample size but decreases with the mean of the sample. These findings and others by Kareev et al. (2002) and Weber, Shafir, and Blais (2004) suggest that the coefficient of variation (the standard deviation divided by the mean) might be a good predictor of intuitive variability judgments based on experience.

Adding to the complexity of the situation is the fact that studies of the in-group heterogeneity effect have used many different measures of perceived group heterogeneity (see Boldry et al., 2007, for a review of the measures used in empirical research). Some studies assessed characteristics of the subjective distribution of the focal trait in the groups, such as range (e.g., Quattrone & Jones, 1980) or variance (e.g., Linville et al., 1989), or the probability of differentiation (e.g., Linville et al., 1989). Other measures relied on the perceived similarity between group members (e.g., Alves, Koch, & Unkelbach, 2016) or the number of subgroups that a participant can generate (e.g., Park, Ryan, & Judd, 1992). Yet others relied on measures of confusion in recall or recognition of the information about the groups (e.g., Ostrom et al., 1993). In this section, we consider all these measures of subjective variability. For each one, we propose a corresponding sample-based measure and we show that our sampling mechanism can lead to the emergence of an in-group heterogeneity effect. We begin with the analysis of estimators of variance different from the ones we analyzed in the previous section. Then we consider estimators of variability that are not specifically estimators of variance.

1.3.1 Other estimators of Variance

Uncorrected Sample Variance

As explained in the introduction, several prior papers on the in-group heterogeneity effect relied on the uncorrected sample variance (Judd & Park, 1988; Park & Judd, 1990; Rubin & Badea, 2007). This is a statistically biased estimator. Linville et al. (1989) proposed that this bias could contribute to explaining the in-group heterogeneity effect. Research in other areas used a similar argument to explain other judgment biases (e.g., Juslin, Winman, & Hansson, 2007; Kareev et al., 2002). The uncorrected sample variance for group g at the end of period t is defined as follows:

$$V^u_{t,g} = \frac{1}{n_{t,g}} \sum_{j=-1}^{t} (x_{j,g} - \bar{x}_{t,g})^2 I_{j,g}. \tag{1.4}$$

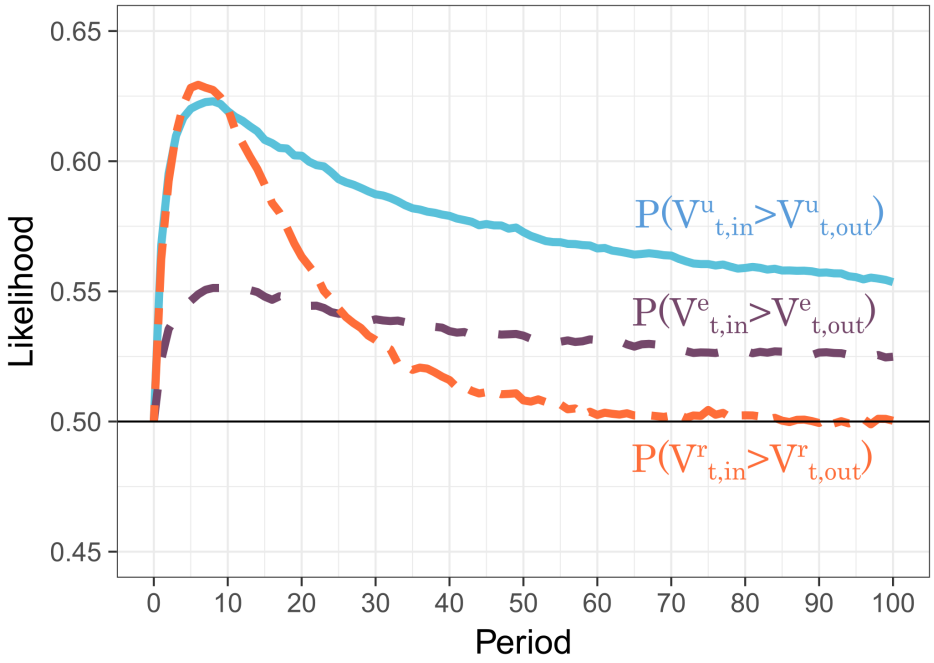


Figure 1.6: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability as a function of time for different measures of variability. The measures include Uncorrected Sample Variance ($P(V^u_{15,in} > V^u_{15,out})$), Variance with Differential Weighting ($P(V^e_{15,in} > V^e_{15,out})$), Recency Weighted Variance ($P(V^r_{15,in} > V^r_{15,out})$). Each point is based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma^2_{in} = \sigma^2_{out} = 1$. The Variance with Differential Weighting is computed with $\alpha = 0.39$. The Recency Weighted Variance is computed with $b_x = 0.3$.

The components of this formula are the same as those in the formula for the corrected sample variance (eq. 1.3). This estimator is biased for small samples, and the size of the bias is stronger the smaller the sample (eq. 1.2). Unsurprisingly, simulations of our model based on this estimator lead to a stronger in-group heterogeneity effect (Figure 1.6). For example, after 15 periods, the likelihood that the estimate of in-group variability is higher than the estimate of out-group variability is $P(V^u_{15,in} > V^u_{15,out}) = .61$. This number was .55 with the corrected sample variance.

Sample Variance with Differential Weighting

A study by Beach and Scopp (1968) proposed that people are more sensitive to smaller deviations than to larger deviations from the mean. They proposed an alternative variance estimator where the deviations are taken not to the power of

two but to another exponent:

$$V^e_{t,g} = \frac{1}{n_{t,g}} \sum_{j=-1}^{t} \mid x_{j,g} - \bar{x}_{t,g} \mid^{\alpha} I_{j,g}, \tag{1.5}$$

where $\alpha > 0$ is a parameter that can be estimated from data. The authors estimated it to be much smaller than 2: 0.39. Such estimator favors smaller deviations and, therefore, will affect the magnitude of the in-group heterogeneity effect. We simulated our model with this value for α. Again, we observe the emergence of an in-group heterogeneity effect (see Figure 1.6). After 15 periods, $P(V^e_{15,in} > V^e_{15,out}) = .55$ (with $r = .75$).

Recency-Weighted Sample Variance

In the analyses we have reported so far, the variance estimators were computed on the basis of the whole set of sampled observations - we assumed perfect memory. The psychological realism of this assumption is questionable. Therefore, here we analyze a model where we assume that the agent stores an estimator of variability and updates it sequentially on the basis of additional information. This approach is similar to models of belief updating commonly used in investigations of attitude formation (e.g. Denrell, 2005; Hogarth & Einhorn, 1992; March, 1996).

Let $V^r_{t,g}$ be the estimate of the variance at the end of the period t and $\hat{x}_{t,g}$ be the estimate of the mean. If the agent samples group g in period t, she updates her estimates as follows:

$$V^r_{t,g} = (1 - b_x)[V^r_{t-1,g} + b_x(x_{t,g} - \hat{x}_{t-1,g})^2], \tag{1.6}$$

and:

$$\hat{x}_{t,g} = b_x \hat{x}_{t-1,g} + (1 - b_x) x_{t,g}, \tag{1.7}$$

where $b_x \in [0,1]$ is the weight of the most recent observation. If the group is not sampled, the estimators of mean and variance do not change. Note that this model implies an exponential memory decay. The strength of the decay increases with the size of the parameter b_x.

Computer simulations show that an in-group heterogeneity effect emerges also in this case (see Figure 1.6).

1.3.2 Other estimators of variability

Coefficient of Variation

In an important paper, Weber et al. (2004) demonstrated that the coefficient of variation (the standard deviation divided by the mean) is a very good predictor of risky choice. Although our focus is not on risky choice, the study of risk behavior and the study of perceived variability are related. This is because the literature

on risky choice defines a risky alternative as an alternative with a variable payoff distribution. The findings from Weber and colleagues suggest that the coefficient of variation could be a relevant measure of perceived variability.

The coefficient of variation is based on the standard deviation. Therefore, biases that affect perceived variance are likely to also affect the coefficient of variation. To see this, let $CV_{t,g}$ denote the coefficient of variation based on observations of the group g until the end of period t. We define it as the ratio of the sample standard deviation over the sample mean:[6]

$$CV_{t,g} = \frac{\sqrt{V^c_{t,g}}}{\bar{x}_{t,g}}. \tag{1.8}$$

We have:

$$P(CV_{t,in} > CV_{t,out}) = P\left(\frac{V^c_{t,in}}{V^c_{t,out}} > \left(\frac{\bar{x}_{t,out}}{\bar{x}_{t,in}}\right)^2\right). \tag{1.9}$$

In our simulations, we assumed that the focal feature had a distribution with mean 0. But it is easy to relax this assumption and simulate a setting where the means for the in-group and the out-group are not 0. Suppose first that the means of the two groups are the same ($\mu_{in} = 1$, $\mu_{out} = 1$) as well as the variance ($\sigma_{in} = 1$, $\sigma_{out} = 1$). In this case, $P(CV_{t,in} > CV_{t,out}) \sim P(V^c_{t,in} > V^c_{t,out})$ and thus the results are very close to those obtained with the sample variance. For example, after 15 periods $P(CV_{15,in} > CV_{15,out}) = .54$ and $P(V^c_{15,in} > V^c_{15,out}) = .55$ (See Figure 1.7).

Now suppose that the mean of the distribution of the focal feature is lower for the in-group than for the out-group ($\bar{x}_{t,in} < \bar{x}_{t,out}$). In this case, the in-group heterogeneity effect is stronger. For example, with $\mu_{in} = 1$ and $\mu_{out} = 1.2$, we have $P(CV_{15,in} > CV_{15,out}) = .65$.

By contrast, when the mean of the distribution of the focal feature is higher for the in-group than for the out-group ($\bar{x}_{t,in} > \bar{x}_{t,out}$), the in-group heterogeneity effect does not always emerge and persist. For example, with $\mu_{in} = 1.1$ and $\mu_{out} = 1$, we have $P(CV_{t,in} > CV_{t,out}) > .5$ until $t = 5$. Then it is lower than .5. For example, after 15 periods, we have $P(CV_{15,in} > CV_{15,out}) = .47$.

If the difference in means is large enough, the in-group heterogeneity does not emerge at all. In fact, an opposite pattern of in-group homogeneity can emerge. For example, with $\mu_{in} = 1.5$ and $\mu_{out} = 1$, we have $P(CV_{t,in} > CV_{t,out}) < .5$ for all $t \geq 1$. We return to the in-group homogeneity effect in a later section.

These analyses tentatively suggest that the in-group heterogeneity effect might be more likely to emerge for undesirable features than for desirable features. This is because the literature on the in-group out-group bias indicates that people are likely to evaluate their in-groups more positively than out-groups. In our model,

[6] We could not find information in the earlier literature whether the relevant sample variance was the uncorrected or the corrected sample variance. Our results are similar in both cases.

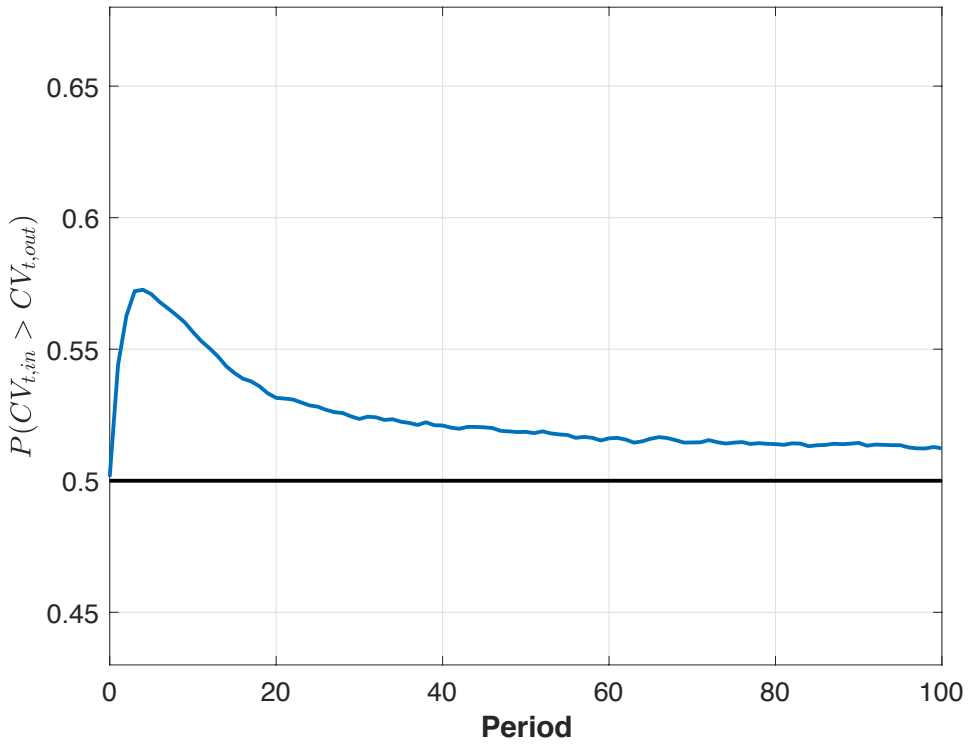


Figure 1.7: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(CV_{t,in} > CV_{t,out})$, when the measure of variability is the coefficient of variation (eq. 1.8). Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 1, \sigma_{in} = 1, \sigma_{out} = 1$.

this translates into assuming that the mean for the in-group is larger than for the out-group when the feature is desirable and that the mean for the in-group is smaller than the mean for the out-group when the feature is undesirable. Whether this prediction makes sense from an empirical standpoint depends on whether the coefficient of variation is a good predictor of perceived group variability in naturally occurring environments.

Similarity

Another frequently used measure of group heterogeneity consists in asking participants to rate the similarity between members of the groups. For example, Boldry and Gaertner (2006, p 389) use the following question: 'to what degree are all members of the group X similar in terms of feature Y' (see also Quattrone and Jones (1980); Badea, Brauer, and Rubin (2012)). One study used a spatial task where participants were asked to position group members on the screen. The similarity was measured as the average distance between the group members (Alves et al., 2016).

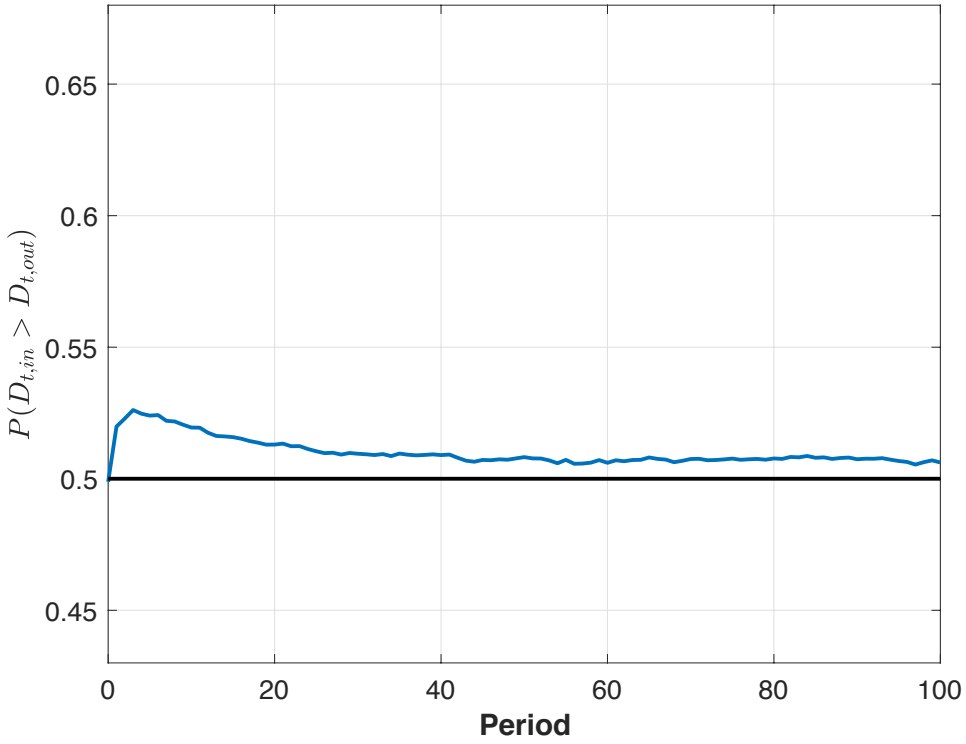


Figure 1.8: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(D_{t,in} > D_{t,out})$, when the variability estimate is the average distance between pairs of group members (eq. 1.10). Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma^2_{in} = \sigma^2_{out} = 1$.

We use a similar method as the latter study and compute the average distance between any two group members. This leads to a measure that is the converse of similarity: the higher this average distance, the lower the average similarity. To keep things simple, we use the absolute value of the difference between the two observations of the focal feature.

$$D_{t,g} = \frac{1}{n_{t,g}(n_{t,g}-1)} \sum_{x_1, x_2 \in O_{t,g}} |x_1 - x_2|. \tag{1.10}$$

An in-group heterogeneity effect emerges after the first period. Figure 1.8 shows that $P(D_{t,in} > D_{t,out}) > .5$ for $t \geq 1$. For example, after 15 periods, we have $P(D_{15,in} > D_{15,out}) = .52$ (with $r = .75$).

Range

A widely used measure of variability is the range spanned by group members on the focal dimension (Quattrone & Jones, 1980; Jones, Wood, & Quattrone, 1981;

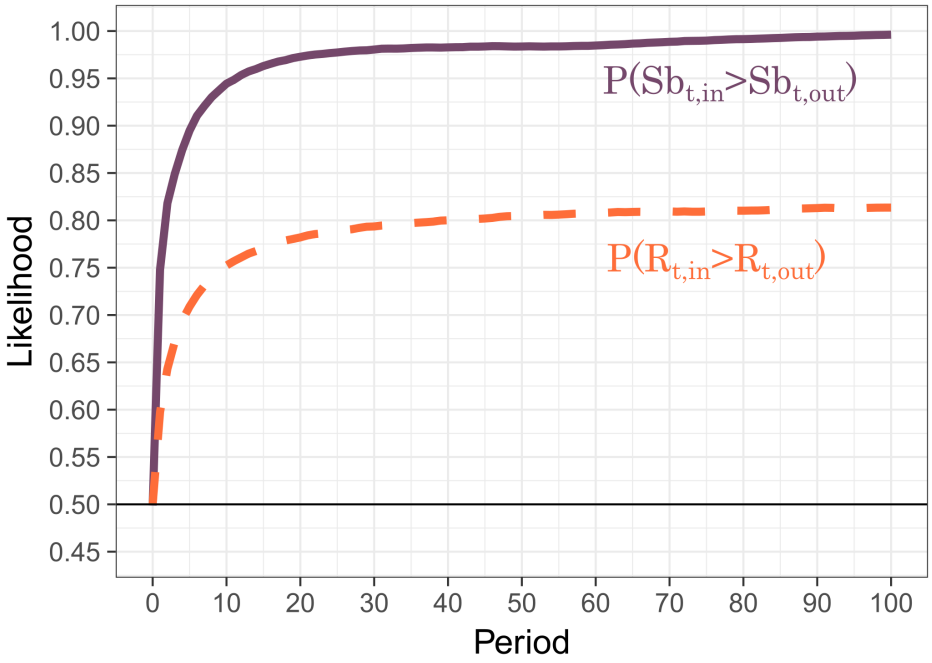


Figure 1.9: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, when the variability estimate is the range, $P(R_{t,in} > R_{t,out})$, and when it is the number of subgroups, $P(Sb_{t,in} > Sb_{t,out} \mid Sb_{t,in} \neq Sb_{t,out})$. Based on 10^5 simulations with $r = .75, \mu_{in} = \mu_{out} = 0, \sigma^2_{in} = \sigma^2_{out} = 1$.

Rubin, Hewstone, & Voci, 2001; Boldry & Gaertner, 2006). Usually, the participants are asked to indicate how extreme a group member can be on either side of the spectrum. Then the low estimate is subtracted from the high estimate. In the context of our model, we can compute the experienced range as the maximum minus the minimum in the sample. Let $O_{t,g}$ denote the set of periods at which the agent sampled group g, until the end of period t.

$$R_{t,g} = \max_{j \in O_{t,g}} x_{j,g} - \min_{j \in O_{t,g}} x_{j,g}, \tag{1.11}$$

An in-group heterogeneity effect emerges after the first period. Figure 1.9 show that $P(R_{t,in} > R_{t,out}) > .5$ for $t \geq 1$. Moreover, this probability is increasing with the number of periods. This is because the range cannot decrease when the sample size increases.

Number of subgroups

In a minority of studies, the participants were asked to generate 'sorts or types' that can describe the groups (Park et al., 1992; Linville, Fischer, & Yoon, 1996).

In the studies that used this measure, an in-group heterogeneity effect emerged. Participants tended to generate more subgroups of in-group than of the out-group.

A variation of our model adapted to this setting produces this result. Research on categorization has shown that people tend to create additional categories when a new observation is far from observations in existing categories. The leading models are very sophisticated (e.g., Anderson, 1991; Sanborn, Griffiths, & Navarro, 2010). Here we provide an illustration using a very simple model that captures the essence of this process. Imagine that both the in-group and the out-group could be divided into up to 10 subgroups. We define boundaries between the subgroups using deciles of the normal distribution (we keep the assumption that the focal feature follows a standard Normal distribution for the two groups). That is, an observation that is in the first decile would be in the first subgroup, an observation in the second decile would be in the second subgroup, etc.

We measure the number of types in a group g by counting the number of subgroups that are being 'hit' by the sample of observations for the group g. We denote it by $Sb_{t,g}$. For example, if all the observations fall in the same decile, group variability is minimal: $Sb_{t,g} = 1$. If, by contrast, the observations span the 10 deciles, group variability is maximal: $Sb_{t,g} = 10$.

With this measure, we do not characterize the in-group heterogeneity effect as $P(Sb_{t,in} > Sb_{t,out}) > .5$ because of its the discrete nature. The probability that the variability of the two groups is the same is positive. Therefore, it could happen that $P(Sb_{t,in} > Sb_{t,out}) < .5$ even though the in-group is likely to be seen as the more variable: $P(Sb_{t,in} > Sb_{t,out}) > P(Sb_{t,in} < Sb_{t,out})$. In such setting, we will say that there is an in-group heterogeneity effect when $P(Sb_{t,in} > Sb_{t,out} \mid Sb_{t,in} \neq Sb_{t,out}) > .5$. [7]

In this case as well, an in-group heterogeneity effect emerges after the first period (see Figure 1.9). For example, after 15 periods, $P(Sb_{15,in} > Sb_{15,out} \mid Sb_{15,in} \neq Sb_{15,out}) = .96$ (with $r = .75$). Additional simulations show that strength of the effect increases with the sampling advantage for the in-group, r. For example, if $r = 0.8$, then $P(Sb_{15,in} > Sb_{15,out} \mid Sb_{15,in} \neq Sb_{15,out}) = .98$.

Probability of Differentiation

In several studies, participants were asked to recreate the distribution of a trait (e.g. 'friendliness') of the members of the groups over a set of 'bins' (Linville et al., 1989; Judd, Ryan, & Park, 1991; Judd & Park, 1988). In these studies, a measure of variability called 'probability of differentiation' was used. It is the probability that two randomly selected members of a group differ on the focal trait. In the original studies, the measure was based on discrete distributions with 5 values (from 1 to

[7] This is formally equivalent to $P(Sb_{t,in} > Sb_{t,out}) > P(Sb_{t,in} < Sb_{t,out})$. In settings where the measure of variability is continuous (denote it $V_{t,g}$), we have $P(V_{t,in} > V_{t,out}) = P(V_{t,in} > V_{t,out} \mid V_{t,in} \neq V_{t,out})$. In discussions of these measures we used the simpler formula $P(V_{t,in} > V_{t,out})$ to make the text easier to read.

5). The probability of differentiation was defined as:

$$Pd = 1 - \sum_{i=1}^{5} p_i^2, \tag{1.12}$$

where p_i is the density of value i according to the elicited distribution ($i \in \{1,2,3,4,5\}$).

It is possible to adapt our model setup to this setting by using discrete (instead of continuous) feature distributions. As an illustration, we will assume the focal feature has 5 levels, with probabilities $(0.1, 0.2, 0.4, 0.2, 0.1)$. This distribution is one of the examples studied in Linville et al. (1989).

Let $Pd_{t,g}$ denote the probability of differentiation for the group g based on the sampled distribution of the focal feature for this group until the end of period t. It is computed using eq. 1.12. We will say there is there is an in-group heterogeneity effect when the in-group is more likely to be perceived as more variable (rather than less variable) than the out-group: $P(Pd_{t,in} > Pd_{t,out} \mid Pd_{t,in} \neq Pd_{t,out}) > .5$ (with $r = .75$).[8]

In this case as well, an in-group heterogeneity effect emerges after the first period. For example, after 15 periods, $P(Pd_{15,in} > Pd_{15,out} \mid Pd_{15,in} \neq Pd_{15,out}) = .7$ (See Figure 1.10). Additional simulations show that strength of this effect increases with the sampling advantage for the in-group, r. For example, if $r = 0.8$, $P(Pd_{15,in} > Pd_{15,out} \mid Pd_{15,in} \neq Pd_{15,out}) = .75$.

Proportion of group members who possess a trait

In several studies, participants were asked to indicate the proportion of group members possessing stereotypical trait (Boldry & Gaertner, 2006; Park & Judd, 1990; Park et al., 1992; Quattrone & Jones, 1980; Ryan, Judd, & Park, 1996). A higher percentage was interpreted as an indication of lower group heterogeneity.

It is possible to analyze this kind of setting with our model by changing the distribution of the focal feature. More specifically, suppose the focal feature is binary (absent or present) and that the probability that a group member possesses it is p_g. We use as a measure of perceived group variability one minus the proportion of observations with the focal feature:

$$Pr_{t,g} = 1 - \frac{1}{n_{t,g}} \sum_{j=-1}^{t} x_{j,g}, \tag{1.13}$$

where $x_{j,g} = 1$ if the focal feature is present and $x_{j,g} = 0$ if it is absent (as before, $n_{t,g}$ is the number of observations of group g until the end of period t.

Suppose the feature is prevalent in the population, but equally stereotypical of the in-group and the out-group: $p_{in} = p_{out} = .85$. An in-group heterogeneity effect emerges after the first period (Figure 1.11). After 15 periods, we have $P(Pr_{15,in} >$

[8] Just as with the number of subgroups, there is often a non-zero probability that the two groups have exactly the same sample variability.

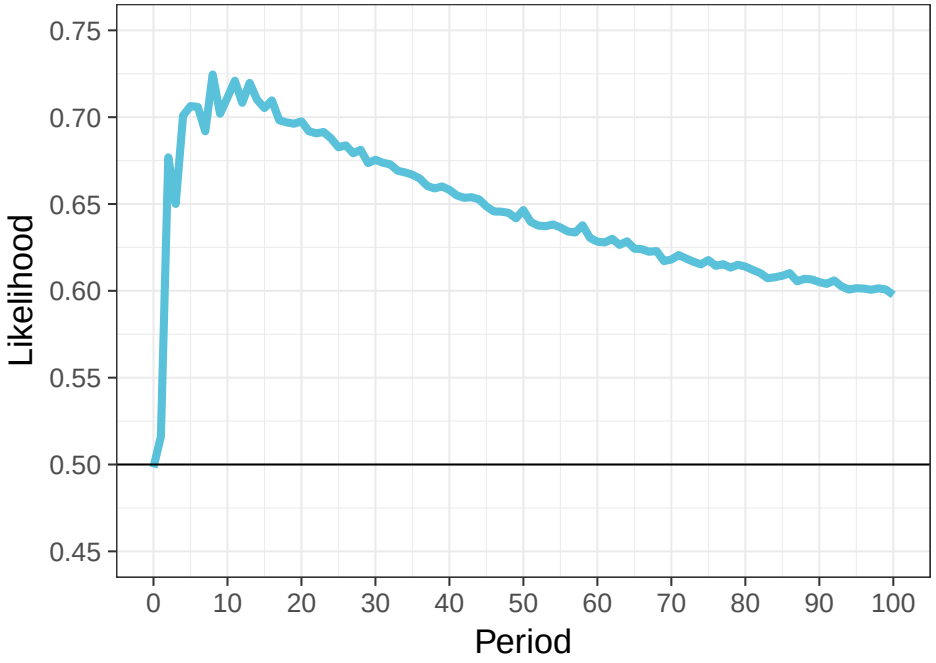


Figure 1.10: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(Pd_{t,in} > Pd_{t,out} | Pd_{t,in} \neq Pd_{t,out})$ when the measure of variability is the probability of differentiation. Each point is based on 10^5 simulations with $r = .75$ and a discrete distribution with 5 levels and the following frequencies: $(0.1, 0.2, 0.4, 0.2, 0.1)$.

$Pr_{15,out} \mid Pr_{15,in} \neq Pr_{15,out}) = .54$ (with $r = .75$). If the trait is rare, the opposite effect emerges: an in-group *homogeneity* effect. Suppose $p_{in} = p_{out} = .15$. After 15 periods, we have $P(Pr_{15,in} > Pr_{15,out} \mid Pr_{15,in} \neq Pr_{15,out}) = .46$ (with $r = .75$). Additional simulations show that strength of this effect increases with the sampling advantage for the in-group, r.

1.3.3 Discussion

In this section, we considered a number of measures of group variability used in the prior empirical and theoretical literatures. For each measure, we proposed a way it could be constructed on the basis of the sampled observations of the group. We showed that our sampling mechanism could produce an in-group heterogeneity effect for all these measures. Still, there exist measures of group heterogeneity that we have not discussed. This is because it is unclear how these measures can be characterized by the sample properties. For example, in several studies, participants were asked to recall or recognize traits of group members (Ostrom et al., 1993; Lorenzi-Cioldi, 1998; Stewart, Vassar, Sanchez, & David, 2000; Ratcliff,

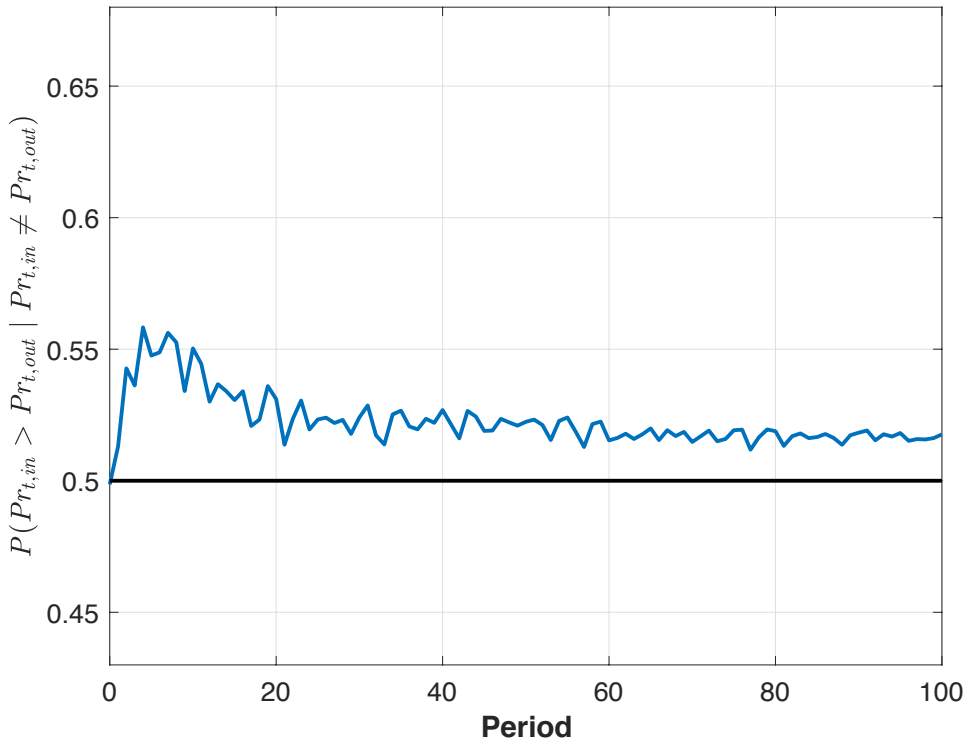


Figure 1.11: Likelihood that the estimate of in-group variability is higher than the estimate of out-group variability, $P(Pr_{t,in} > Pr_{t,out} \mid Pr_{t,in} \neq Pr_{t,out})$, when the measure of variability is the perceived proportion of group members that possess a trait (eq. 1.13). Based on 10^5 simulations with $r = .75$, and $p_{in} = p_{out} = .85$.

Hugenberg, Shriver, & Bernstein, 2011). Higher confusion among group members was interpreted as lower perceived group heterogeneity. It is unclear how our sampling mechanism could explain the findings of these studies without invoking specific assumptions about the storage and retrieval of information in memory.

All of our analyses relied on an assumption that the variability of the sample drives perceived group heterogeneity. We provide evidence for this assumption in a later section. Next, we discuss how the predictions of our model relate to the existing literature on the in-group heterogeneity effect.

1.4 Relation to the Existing Literature

Most prior explanations of the in-group heterogeneity effect invoke differences in how information about in-group and out-group is processed. Here we discuss how our explanation differs from this prior work. We use a taxonomy similar to Ostrom and Sedikides (1992).

Several explanations rely on motivated cognition (Kunda, 1990). The first one invokes people's desires for positive identities. Those who want a positive social

identity are motivated to view their in-groups more positively than other groups (Tajfel, 1982). At the same time, heterogeneity is frequently perceived as a positive feature of social groups (Ostrom & Sedikides, 1992). Therefore, people are motivated to perceive the in-group as more heterogeneous than out-groups. A related explanation invokes people's desire for distinct identities. A more heterogeneous in-group allows people to see themselves as unique within the in-group. Thus, people are motivated to see their in-groups as heterogeneous (Pickett & Brewer, 2001). Yet another explanation based on motivated cognition notes that it is easier to dehumanize more homogeneous groups (N. Haslam, 2006; Brewer, 1999). Therefore, if the out-group is perceived as less variable than the in-group, it is easier to justify negative attitudes and even cruel actions towards out-group members.

The second type of explanation notes that people tend to have prior beliefs that the out-group is more homogeneous. Park and Hastie (1987) showed that if participants first observed exemplars from a group followed by a description of its general characteristics, they perceived this group as more variable compared to when they observed that information in reversed order. This suggests that the prior about homogeneity affects how information is encoded. This finding implies an in-group heterogeneity effect under a (reasonable) assumption that people often learn descriptions of out-groups before interacting with some of their members (e.g. through stereotypes communicated by others in their environment) whereas they learn about in-groups by direct observations.

The third type of explanation notes that the self is part of the in-group (Park & Judd, 1990). Since the self is often perceived as particularly differentiated and unique, this would contribute to an impression that the in-group is more heterogeneous than the out-group.

The fourth type of explanations suggests that information about different groups is encoded and retrieved in different fashions. For example, Ostrom et al. (1993) found that information about in-group members is stored in categories related to individual information whereas the information about the out-group members is stored in categories related to stereotypical attributes. Therefore, when the information is recalled, the in-group tends to be associated with more individuating information compared to stereotype based homogeneous information about the out-group. In terms of recall, Park and Judd (1990) suggested that participants recall more extreme exemplars about in-groups than about out-groups. This suggests that memory search processes might differ across in-group and out-group.

These four types of explanation emphasize features of information processing. By contrast, our explanation focuses on properties of the sample of information on which the mind operates. Because the two classes of explanations focus on different levels (information sampling and processing), they do not contradict each other. Rather, our sampling explanation complements the explanations that focus on information processing. Our analyses and the experimental findings discussed above suggest that both types of mechanisms likely play a role in explaining the in-group heterogeneity effect.

As explained in the introduction, Linville et al. (1989) also proposed a sampling-

based explanation for the in-group heterogeneity effect. They characterized the effect in terms of the mean of the variability estimators: $E(V_{in}) > E(V_{out})$. Our characterization differs. It is in terms of the probability that an individual will see the in-group as more variable than the out-group: $P(V_{t,in} > V_{t,out}) > .5$. The characterization used by Linville et al. (1989) has a much narrower scope than ours because it only works if we assume that people use a statistically biased estimator of variability. By contrast, our characterization works even if we relax this assumption. Our analyses build on and extend the idea introduced by Linville et al. (1989) and show that a sampling-based mechanism can produce an in-group heterogeneity effect for most measures of heterogeneity used in the prior literature. This implies that a sampling explanation could contribute to explaining many of the empirical findings in the prior literature. It is the case in particular for studies that assessed participants' perceived group heterogeneity for social groups in their naturally occurring environments. By contrast, a sampling explanation cannot explain findings based on a minimal group paradigm (Boldry & Gaertner, 2006; Rubin & Badea, 2007, 2010), because there is no sampling asymmetry in this case.

1.5 In-group Homogeneity Effect

Even though most studies on the perception of group heterogeneity documented an in-group heterogeneity effect, several papers have documented an opposite pattern: an in-group *homogeneity* effect (e.g., B. Simon & Pettigrew, 1990; Rubin & Badea, 2007). Here we show that these findings are compatible with a sampling explanation.

1.5.1 Difference in True Group Variabilities

Like the in-group heterogeneity effect, the in-group homogeneity effect might be a reflection of a difference between true group variabilities. The in-group might be more homogeneous than the out-group on a feature that is stereotypical for the in-group. For example, in Experiment 1 by S. A. Haslam, Oakes, Turner, and McGarty (1995), Australians indicated from a list of words traits that "seem the most typical of people" from Australia and the US. Participants were asked to choose 5 of these traits and to provide the percentages of people that possess that characteristic in each country. The 5 traits chosen for Americans were regarded as stereotypical of Americans and counter-stereotypical of Australians. Similarly, the traits chosen as stereotypical of Australians were considered as counter-stereotypical of Americans. Note that these definitions are specific to each participant, as every participant independently indicated which traits they regarded as stereotypical of each nation.

Consider a trait listed as typical of Australians. Participants indicated their estimated percentage of Australians and Americans with this trait. The stated percentages of Australians were higher than the stated percentages of Americans. In

other words, Australians perceived their in-groups as more homogeneous than their out-groups on stereotypical traits (assuming that group homogeneity is defined as the proportion of group members who possess a given trait).

To capture this, we can use a variation of our model with the 'proportion of group members who possess a trait' (analyzed in the previous section). Suppose that the feature is binary. 85% of the in-group members possess it whereas 50% of the out-group members possess it ($p_{in} = .85; p_{out} = .5$). The in-group is less variable than the out-group. In this case, an in-group homogeneity effect will emerge even if there is a sampling advantage for the in-group. For example, with $r = .75$, there is an in-group homogeneity effect in all periods: $P(Pr_{t,in} > Pr_{t,out} \mid Pr_{t,in} \neq Pr_{t,out}) < .5$. After 15 periods, this probability is .08. A similar pattern occurs if there is no sampling asymmetry ($r = .5$). Unsurprisingly, a difference in true variabilities is enough to imply a similar difference in sample variabilities.

It is also possible to capture settings where the focal feature is stereotypical of the in-group with our baseline model (with continuous features). In this case, it is enough to assume that the true variability of the in-group is lower than the true-variability of the out-group: $\sigma^2_{in} < \sigma^2_{out}$. We discussed this setting in the sensitivity analyses reported in the 'Simple Model' section. In this case, the variabilities of the two groups will tend to be underestimated. More importantly, the in-group variability will be underestimated to a lower extent than the out-group variability. If the difference in the extent of underestimation is smaller than the difference in true variabilities, our model implies the emergence of an in-group *homogeneity* effect, in line with the true difference in variabilities. But if the difference in true variabilities is small, our model can lead to the emergence of an in-group *heterogeneity* effect (see Figure 1.4).

1.5.2 Out-group sampling advantage

Finally, there exists evidence that when the in-group is a minority it tends to be judged as more homogeneous than the out-group (B. Simon & Pettigrew, 1990; Voci, Hewstone, Crisp, & Rubin, 2008). There is also evidence that minority members tend to be frequently exposed to majority members, especially when the minority is small. For example, in the GSS survey data analyzed by Marsden (1987), hispanic and black respondents listed more members of different racial groups among their closest contacts as compared to white respondents. This suggests that minority members interact more frequently than majority members with out-group members. At the extreme, minority members might interact more frequently with majority members than with members of their own minority group. It is possible to capture this kind of situation by assuming $r < .5$ in our model. As shown on Figure 1.3, our model predicts the emergence of an in-group homogeneity effect. When the in-group is a minority, it might be both less variable than the out-group (it is smaller) and there is a sampling advantage for the out-group. In this case, two factors that contribute to the in-group homogeneity effect operate simultaneously. An in-group homogeneity effect is thus all the more likely to

emerge.

1.6 The Role of Group Size

In the baseline model, we have implicitly assumed that groups are of infinite size. In reality, however, groups often have different (finite) sizes. The difference in group sizes has implications for the amplitude of the in-group heterogeneity effect. This is because it might affect both the relative sampling probabilities and the true group variabilities.

1.6.1 Model Description

The model is very similar to the simple model. But instead of assuming two groups of infinite size, we assume the two groups have finite sizes N_{in} and N_{out}. This poses an additional constraint on the sampling behavior of the agent. She is less likely to sample from the smaller group.

Distribution of the Focal Feature

The focal feature for both groups follows a normal distribution with mean 0 and variance 1: $\mu_{in} = \mu_{out} = 0$, $\sigma^2_{in} = \sigma^2_{out} = 1$. The in-group is a set of N_{in} independent draws from the $N(0,1)$ distribution. Similarly, the out-group is a set of N_{out} independent draws from the $N(0,1)$ distribution.

Sampling Rule

In each period, the agent samples from the in-group or from the out-group (without replacement). The probability the agent samples a particular member of the in-group is proportional to r. The probability the agent samples a particular member of the out-group is proportional to $1-r$. This implies that the probability that the agent samples from the in-group depends on r and the number of group members that have not been sampled yet in both groups. Let $k_{t,g}$ denote the number of members sampled from group g by the end of period t:

$$k_{t,g} = \sum_{j=-1}^{t} I_{j,g}$$

where $I_{j,g}$ is an indicator variable equal 1 if group g is sampled in period j (and equal to 0 otherwise). The probability the agent samples the in-group in period t is:

$$P_{t+1,in} = \frac{r(N_{in} - k_{t,in})}{r(N_{in} - k_{t,in}) + (1-r)(N_{out} - k_{t,out})} \tag{1.14}$$

This probability decreases with $k_{t,in}$ and increases with $k_{t,out}$. That is, for all periods after t such that $k_{t,in} = N_{in}$ the probability becomes 0. At the same time for all periods after t such that $k_{t,out} = N_{out}$ the probability becomes 1.

Perceived Group Variability

We assume that perceived group variability is given by the corrected sample variance (eq. 1.3). Note here that because the groups are finite the real variability of group g is the corrected sample variance of the sample of size N_g.

Analysis

First, let us consider the case where the out-group is smaller than the in-group. The model leads to a stronger in-group heterogeneity effect than the baseline model. Suppose $r = .75, N_{in} = 50, N_{out} = 10$. After 15 periods, we have $P(V^c_{15,in} > V^c_{15,out}) = .62$. This quantity was .55 with the baseline model. In this case, the fact that the out-group is smaller decreases the probability that the agent would sample from it. This is because the difference in group sizes has two parallel effects that reinforce each other: the agent is more likely to sample members of the in-group and the in-group is likely to be more variable than the out-group (because it is larger). The later effect is not present in the baseline model which explains the stronger magnitude of the effect in the present simulations.

Second, suppose the out-group is larger than the in-group. For illustration let us assume $N_{in} = 10$ and $N_{out} = 50$. In this case, the model predictions depend on the sampling advantage of the in-group members. For most values of r, an in-group homogeneity effect will emerge. For example, with $r = .75$ after 15 periods $P(V^c_{15,in} > V^c_{15,out}) = .47 < .5$. But for very high values of r, the sampling advantage compensates for the asymmetry in group sizes. For example, with $r = .9$, we get $P(V^c_{15,in} > V^c_{15,out}) = .5$.

There are many possible implementations of the model with finite group sizes. Although the amplitude of the effect will differ depending on specific assumptions, we believe that an in-group heterogeneity effect will emerge whenever the model setup implies that agents tend to gather larger samples of information about the in-group than about the out-group.

1.7 Sampling from Memory

In the model we analyzed above, we assumed that the agent had perfect memory: the agent used all the sampled information to form variability estimates. In most of our analyses, all observations were given equal weights, with the exception of the model with a recency bias (see above the subsection 'recency-weighted sample variance'). There exists evidence that people's variability estimates might be constructed from samples of observations retrieved online from memory (e.g., Juslin et al., 2007; Kareev et al., 2002). Here, we analyze what happens when we incorporate such sampling from memory in our model.

Distribution of the Focal Feature

There are two groups ($g = in, out$) and the focal feature for both groups follows a normal distribution with mean 0 and variance 1: $\mu_{in} = \mu_{out} = 0$, $\sigma^2_{in} = \sigma^2_{out} = 1$.

Sampling Mechanism

We assume that sampling occurs at two stages:

1. *Sampling from the environment*. In each period, the agent samples one group or the other. The probability the agent samples the in-group in period t is r.

2. *Sampling from memory*. The variability estimates are constructed from a sample of (at most) N observation from the set of observations collected in prior periods. We assume that each observation is equally to be retrieved (independently of the group to which it belongs)

Perceived Group Variability

The perceived group variability for the group g is given by the corrected sample variance of the sample retrieved from memory (eq. 1.3). The difference with our baseline model is that the information sample used to form a group variability estimate is a subset of at most N observations of the set of observations collected about that group.

Analysis

The samples from memory are based on the samples collected from the environment. Therefore, they reflect, to some extent, the asymmetry in the sizes of the samples collected about the two groups. This implies that an in-group heterogeneity effect will emerge when the in-group has a sampling advantage ($r > .5$).

Suppose that $N = 7$. We make this assumption because it corresponds to the number of chunks of familiar information that can be stored in working memory (Miller, 1956). We assume $r = .75$. After 15 periods the amplitude of the in-group heterogeneity effect is close to what was obtained with the baseline model: $P(V^c_{15,in} > V^c_{15,out}) = .56$ (this quantity was .55 with the baseline model).

The predictions of the two models differ when the number of periods is large, however. Suppose that there are 100 periods. With the limited memory model, the size of the effect is the same as with 15 periods: $P(V^c_{100,in} > V^c_{100,out}) = .56$. But with the baseline model, the size of the effect is smaller than what it was with 15 periods: $P(V^c_{100,in} > V^c_{100,out}) = .52$.

The difference in sample variabilities (from the environment) decreases as the sizes of the samples collected about the in-group and the out-group increase. With limited memory, the samples used to form variability estimates are both small, and the sample retrieved about the in-group is frequently larger than the sample retrieved about the out-group. In fact, under the memory sampling mechanism,

the size of the asymmetry in perceived variability remains the same for all periods after period N (because the size of the sample retrieved about the in-group is given by a binomial distribution with parameters .75 and N). More formally, for $t \geq N$, $P(V^c_{t,in} > V^c_{t,out}) = P(V^c_{N,in} > V^c_{N,out})$. The amplitude of the effect thus does not go down as the number of periods becomes large.

It is possible to make different assumptions regarding memory sampling. Suppose instead that the agent samples from memory at most N observations from the in-group and at most N observations from the out-group. In this case, if t is large enough, the agent is likely to have sampled at least N observations from each group (from the environment). The sizes of the retrieved memory samples will be the same: N in both cases. In this case, an in-group heterogeneity effect will not occur: $P(V^c_{t,in} > V^c_{t,out}) \approx .5$.

Existing research on overconfidence has shown that different types of questions trigger different memory sampling processes (Juslin et al., 2007). This suggests that the format of the questions used to elicit perceived group variability might affect the size of the observed asymmetry. We conjecture that questions that ask to compare the variabilities of two groups will trigger the first memory sampling mechanism out of the two discussed above (N observations of which some will be from the in-group and some from the out-group). In this case, an in-group heterogeneity effect is likely to be observed. By contrast, questions that ask for the perceived variability of the two groups separately are likely to trigger the second memory sampling mechanism (N observations from the in-group and N observations from the out-group). In this case, an in-group heterogeneity effect is less likely to emerge (assuming that the true variabilities of the two groups are the same).

Several studies suggested that memory processes could explain the in-group heterogeneity effect (Ostrom et al., 1993; Park & Judd, 1990). They proposed that there are systematic differences in how the information about in-group and out-group is retrieved. By contrast, our analysis suggests that even if there is no difference in retrieval processes, a limited working memory capacity could contribute to explaining the in-group heterogeneity effect.

1.8 Hedonic Sampling

The model we have analyzed so far assumed that sampling behavior was entirely driven by the structure of the environment (i.e., by an exogenous and fixed parameter r). We made this assumption to keep our model as simple as possible. Yet, a large amount of research has shown that prior experiences affect people's propensity to interact and thus sample social groups (for a review, see Denrell, 2005). Here, we show that our results still hold when prior experiences affect sampling behavior.

1.8.1 Model Description

The model is similar to our baseline model, but the agent observes two dimensions of the group when she samples it: an attitudinal dimension, A and another dimension X. We are interested in the perceived variability on the X dimension. Dimension A affects sampling behavior.

Distribution of the Observed Values Rule

For simplicity, we assume the two groups are similar: X and A have the same distribution in the two groups. We assume that X follows a Normal distribution with mean μ_x and variance σ_X: $\mu_{X,in} = \mu_{X,out} = \mu_X$, $\sigma^2_{X,in} = \sigma^2_{X,out} = \sigma_X$. Similarly, we assume that A follows a Normal distribution with mean μ_A and variance σ_A: $\mu_{A,in} = \mu_{A,out} = \mu_A$, $\sigma^2_{A,in} = \sigma^2_{A,out} = \sigma^2_A$.

Belief Updating

Let $A_{t,g}$ denote attitude of the agent toward group g at the end of period t. If she samples group g in period t, two things happen:

1. She updates her attitude toward the group. Her new attitude is a weighted average of her previous attitude and the new observation $a_{t,g}$:

$$A_{t,g} = (1-b)A_{t-1,g} + ba_{t,g}, \tag{1.15}$$

   where $b \in [0,1]$ and $a_{t,g} \sim N(\mu_A, \sigma_A)$. This attitude updating rule has been found to provide good fit to experimental data on sequential choice under uncertainty (see Denrell, 2005 for a review).

2. She obtains an observation $x_{t,g}$ of the non-attitudinal dimension. $x_{t,g} \sim N(\mu_X, \sigma_X)$.

Sampling Rule

To ensure that variability estimates exist for both groups, we assume that the agent has sampled 2 observations from each group before the first period. In the subsequent periods, the sampling rule follows that used in Denrell (2005). In each period, the agent samples the in-group or the out-group based on the current attitude towards that group. The probability that the agent samples the in-group is given by the exponential version of the Luce choice rule:

$$P_{t+1,in} = l + (1-l)\frac{e^{sA_{t,in}}}{e^{sA_{t,in}} + e^{sA_{t,out}}}, \tag{1.16}$$

where s is a parameter that regulates the sensitivity of the sampling probability to the current attitude, and $l \in [0,1]$ is a parameter that corresponds to the sampling advantage of the in-group. The higher l is, the higher is the baseline probability

that the agent will sample the in-group. When l is close to 1, the agent is likely to frequently sample the in-group even if she has a negative attitude toward it ($A_{t,in}$ is low). When l is close to 0, information sampling is mostly driven by the agent attitudes towards the two groups.

Note that this sampling rule does not depend on observations of X. It is easy to relax this assumption, but we wanted to keep the two variables separated to help clarity about the dynamics of the model. It is possible 'collapse' A and X. In this case, the task environment is essentially a bandit problem and analyses of the model lead lead to systematic predictions about the association between perceived risk and reward. This topic is quite distinct from group perception, and thus we leave it for a further research.

Perceived Group Variability

We assume that perceived heterogeneity is the corrected sample variance, as in the first model we analyzed (eq. 1.3).[9]

Model Summary

As before, the agent is subject to an environmental constraint that makes her sample the in-group more frequently than the out-group. But her attitudes toward the groups also affect her sampling behavior: she is more likely to sample again a group with which she has had positive experiences and to avoid a group with which she has had negative experiences.

1.8.2 Analysis

The results are very similar to those obtained with the model without attitudes. Suppose first that the means are the same ($\mu_{X,in} = \mu_{X,out} = 0$ and $b = 0.5, s = 3, l = 0.5$). After 15 periods we have: $P(V^c_{15,in} > V^c_{15,out}) = .56$. This likelihood was .55 with the baseline model (which is equivalent to $s = 0$ and $r = .5 + .5l$). If the means are different (e.g. $\mu_{in} = 0.2$ and $\mu_{out} = 0$ with b, s, l the same as above) the effect is stronger. After 15 periods, we have $P(V^c_{15,in} > V^c_{15,out}) = .57$ (see Figure 1.12).

The amplitude of the effect depends on b, s and l. Higher reliance on the most recent information (higher b) leads to a weaker in-group heterogeneity effect. At the same time, higher sensitivity to the current estimate (higher s) or larger sampling advantage of the in-group (higher l) magnify the effect (see Figure 1.13).

[9] It might seem surprising that we assume recency weighting for attitude formation (A) but not for perceived variability (based on observations of X). We assumed that perceive variability was based on unweighted sample variance to facilitate direct comparison to the baseline model analyzed above. If we define perceived variability as recency-weighted sample variance (see eq. 1.6), similar results hold.

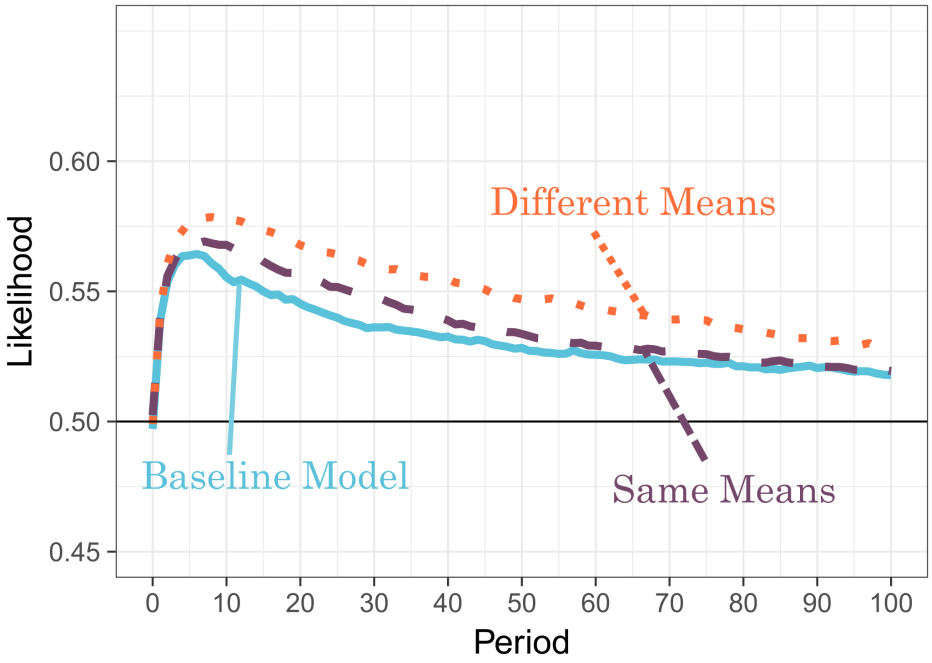


Figure 1.12: Likelihoods that the estimate of in-group variability is higher than the estimate of out-group variability under *hedonic* sampling, $P(V^c_{t,in} > V^c_{t,out})$, when the estimator of variability is the corrected sample variance (eq. 1.3). The likelihoods are estimated for three sets of parameters: '*Baseline Model*' $b = 0.5, s = 0, \mu_{X,in} = \mu_{X,out} = 0$ (solid line); '*Same Means*' $b = 0.5, s = 3, \mu_{X,in} = \mu_{X,out} = 0$ (dashed line); '*Different Means*' $b = 0.5, s = 3, \mu_{X,in} = 0.2, \mu_{X,out} = 0$ (dotted line). Based on 10^5 simulations with $l = 0.5, \mu_A = 0, \sigma^2_A = \sigma^2_{X,in} = \sigma^2_{X,out} = 1$.

1.9 Sample Variability, Sample Size and Perceived Variability

All our analyses rely on the assumption that perceived group variability depends on the variability of the sample obtained about the group. Because this assumption is so crucial to our argument, we now turn to providing evidence in support for its realism. Few studies focused on this relation in the context of the in-group heterogeneity effect other than Linville et al. (1989). But several studies have focused on the relationship between sample variability and perceived variability of the underlying distribution in other settings (e.g., Kareev et al., 2002; Weber et al., 2004). We also identified several data sources (an experiment and two survey data sets) that allow for the measurement of the association between sample variability and perceived variability of the underlying distribution. Below, we report analyses of these data.

We also review existing evidence that supports our claim that perceived vari-

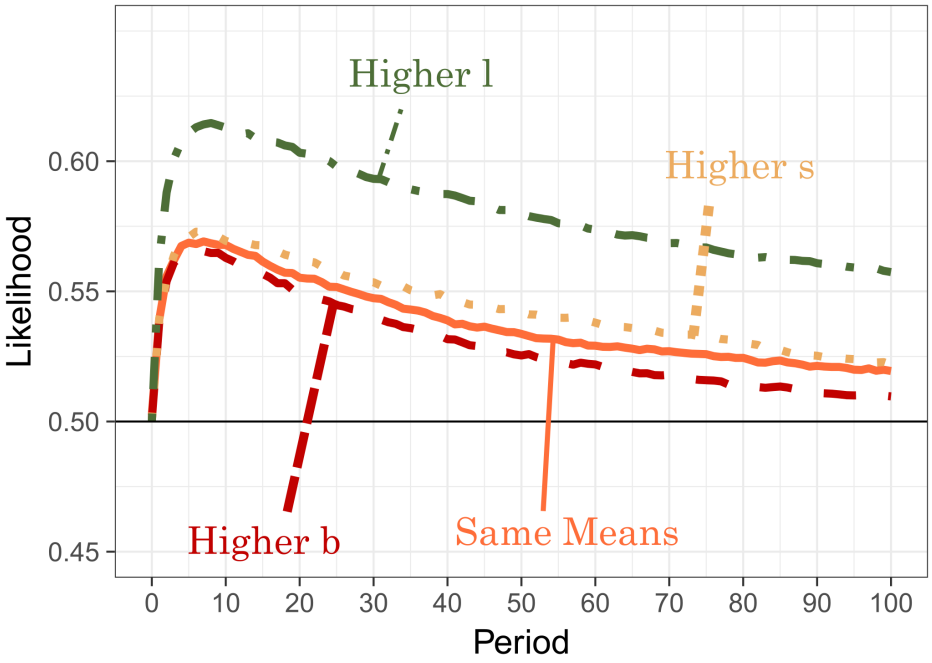


Figure 1.13: Likelihoods that the estimate of in-group variability is higher than the estimate of out-group variability under *hedonic* sampling, $P(V^c_{t,in} > V^c_{t,out})$, when the estimator of variability is the corrected sample variance (eq. 1.3). The likelihoods are estimated for four sets of parameters: '*Same Means*' $b = 0.5, s = 3, l = 0.5$ (solid line);'*Higher b*' $b = 0.8, s = 3, l = 0.5$ (dashed line); '*Higher s*' $b = 0.5, s = 8, l = 0.5$ (dotted line); '*Higher l*' $b = 0.5, s = 3, l = 0.8$ (dot dashed line). Based on 10^5 simulations with $\mu_A = 0, \mu_{X,in} = \mu_{X,out} = 0, \sigma^2_A = \sigma^2_{X,in} = \sigma^2_{X,out} = 1$.

ability increases with sample size. Finally, we report a new experiment that directly tests a major prediction of our model: that when people have two samples of different sizes from the same distribution, most of them will believe that the larger sample comes from a more variable distribution than the smaller sample.

1.9.1 Sample Variability and Perceived Variability

Existing Evidence

Experiment 1 in Kareev et al. (2002) was designed to study how information sampled about an alternative affects its perceived variability. In this experiment, participants went through two tasks with the following structure (here, we describe just one of the tasks): They first observed a population of 28 items that differed from each other on just one dimension. The items were paper cylinders of the same shapes colored up to a certain height. The height of coloring was the focal dimen-

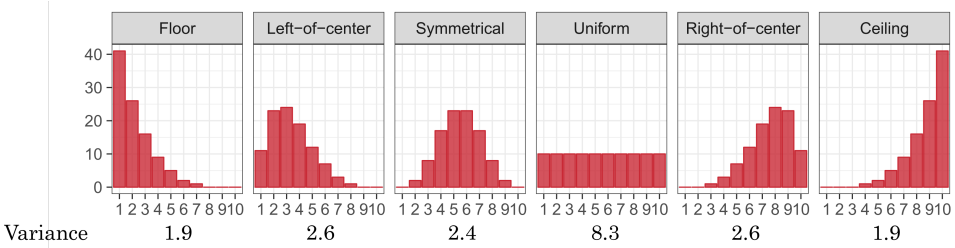


Figure 1.14: Distributions used in Goldstein and Rothschild (2014).

sion. The coloring height was normally distributed with mean 6 cm and standard deviation 1.955 cm.

Participants completed a comparison task: They were shown two additional populations of 28 items and were asked to identify the population most similar to the original. Unbeknown to the participants, one of the two comparison populations was the same as the population they saw. The other comparison population had higher variability (the distribution of coloring height had mean 6 cm and standard deviation 2.112 cm) or lower variability (same mean and standard deviation 1.811 cm).

The authors were interested in the proportion of participants who would select the non-identical population when it had higher or lower variability (the correct choice was to select the identical population). They found that when this alternative had lower variability, it was more likely to be selected than when it had higher variability. This result indicates that participants were sensitive to the variability of the population and that they had a systematic tendency to underestimate the variability. The reason for this underestimation is not limited sampling from the alternatives, as in our baseline model (the participants observed the whole population). Kareev et al. (2002) explain this by noting that people have limited working memory capacity (and provide evidence for this explanation in a subsequent experiment). This explanation is consistent with our simulation of the limited memory model reported above. Nevertheless, this experiment provides evidence that sampling variability affects perceived variability. Another experiment in that paper provides additional evidence. We discuss it in the next subsection.

Experiment 1 in Weber et al. (2004) also provides evidence that people are sensitive to sampled variability. They focused on risky choice situations where one of the alternatives had a sure payoff x and the risky alternative had a probability p to yield a high payoff $y > x$ and $1 - p$ to yield a low payoff. Participants made choices based on experience: they were not provided with a description of the payoff distributions, but instead had to learn these by sampling the two alternatives. The authors found that people were less likely to select the risky alternative when the coefficient of variation (CV) of the risky alternative was high. This indicates that the perceived variability of the risky alternative was influenced by the sampled variability of that alternative. This study thus provides support for our assumption

that perceived variability is affected by sample variability.

An important finding of this study, for the purpose of the original paper, is that the coefficient of variation is a better predictor of risky choice than the variance of the payoff distribution. This finding suggests that in a setting where a decision maker has to decide whether to interact with a member of group A or group B, and the outcome of the interaction is hedonically relevant, then the coefficient of variation might be the most relevant measure of perceived group variability.

Additional evidence comes from the many studies that relied on the 'sampling paradigm' used to study the 'decision-experience gap' (Hertwig, Barron, Weber, & Erev, 2004). In the baseline version of the task, a decision maker faces two alternatives with unknown payoff distributions. One alternative has binary outcomes and unknown probability of success. The other has a sure outcome. She can freely sample the two alternatives for as many periods as she likes. Then she has to choose one of the two alternatives (and gets rewarded according to the payoff drawn from the selected alternative). In this task, researchers have repeatedly found that people make choices as if underweighting the probability of rare events. This behavioral pattern has been (partly) attributed to properties of the sampled outcomes: When the probability of success of the uncertain alternative is small, most decision makers sample success less frequently than expected by the probability of success (Fox & Hadar, 2006; Hertwig et al., 2004; Ungemach, Chater, & Stewart, 2009). This is because the payoff distribution is skewed and sample sizes are small. In summary, research on the sampling paradigm has found that risky choice was influenced by sample variability. This suggests that perceived variability was influenced by sample variability.

New Analysis of Existing Experimental Data

We analyzed data collected by Goldstein and Rothschild (2014). In this online experiment, the authors told the participants that they had a very large bag with balls, that each ball had a number written on it, and that the range of numbers was 1 to 10. Participants were then shown 100 balls from the urn in a random order. It is important to note that the composition of the sample of 100 balls was not 'random', but was generated so as to be as close as possible to the generating distribution. In particular, the sample variance was essentially the same as the variance of the generating distribution. After seeing the sample, beliefs about the distribution of numbers were elicited using a tool designed by the authors called the 'distribution' builder. They also elicited the perceived 10%-90% range.[10] Goldstein and Rothschild (2014) write

> After observing all 100 numbers from a randomly-assigned distribution and shuffle combination, respondents are told "Now imagine we

[10]The distribution and range tasks were two of the tasks they used, they also similarly measured percentiles of the distribution. We focus on the data from these two tasks since it is the most comprehensive assessment of the distributional beliefs of the participants.

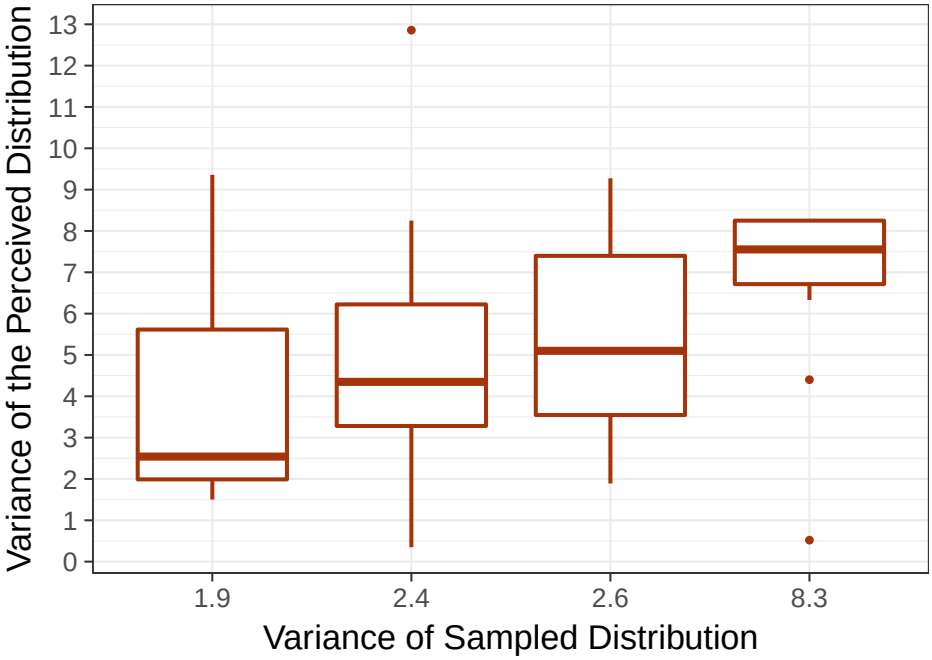


Figure 1.15: Analysis of the Goldstein and Rothschild (2014) data: Box plot of the impact of the higher variance of the sampled distribution on the variance of the perceived distribution.

> throw the 100 balls you just saw back into the bag and mix them up. After that, we draw again 100 balls at random" [...] Respondents [were] asked "How many balls of each value (from 1 to 10) do you think we would draw?" By clicking on buttons beneath columns corresponding to the values from 1 to 10, respondents place 100 virtual balls in ten bins, ultimately creating a 100-unit histogram that should reflect their beliefs about a new sample drawn from the same population that gave rise to the sample they initially observed.

For the range task, participants were told that instead of 100 balls they would just draw one ball. Then they completed the following statements: "I am 90% certain the value of this ball would be greater than or equal to ..." and "I am 90% certain the value of this ball would be less than or equal to ...". In a between participant design, the authors used six distributions (Figure 1.14) which had four unique variances.

Figure 1.15 shows that the variance of the elicited distribution is increasing in the variance of the sampled distribution. A one-way ANOVA shows that this effect is strongly significant ($F(3,117) = 6.76$, $p < 0.001$). Similar but weaker results hold for the range task ($F(3,116) = 2.50$, $p = 0.063$). These results indicate that perceived variability strongly depends on the variability of the sample. Next, we

show that similar results hold in non-experimental settings.

Analysis of LISS Panel Data

We analyzed data from the Longitudinal Internet Studies for the Social Sciences (LISS) Panel. These data are from a representative sample of the Dutch population and were collected by CentERdata in collaboration with Galesic et al. (2012) for a project that explored the relationship between the social circles of the respondents (the individuals with whom they interact most frequently) and their perceptions of the national population as a whole. In this study, the authors asked respondents about 10 characteristics related to their financial situation, friendships, health, work stress and education. The respondents reported their beliefs about the distribution of these characteristics on a 7-point scale. They were also asked to estimate the distribution in the general population of the Netherlands with questions such as "What percentage of adults living in The Netherlands fall into the following categories". In a second wave, participants were asked to provide the distribution in their social circle with questions such as "What percentage of your social contacts fall into the following categories?". 'Social contacts' were defined as "all adults you were in personal, face-to-face contact with at least twice this year." (quoted from the codebook of the second wave of the study).

The authors were interested in how social sampling impacts beliefs about population characteristics. In their analyses, they assumed their available samples about the population were made of their social circles. Here, we rely on the same assumption.

We focus on one specific aspect of the social circle and perceived population distributions: their variance. For each of the 10 characteristics, we regressed the variance of the perceived population distribution on the variance of the social circle distribution. The slope coefficient is positive for all 10 characteristics. It is also positive in a regression that pools the data about all 10 characteristics and includes characteristics fixed effects (coefficient = 0.15, see Table 1.1). It is worth noting that the coefficients are somewhat far from 1. This indicates that the distribution in the social circle is not the only factor affecting the perceived population distribution. This is not surprising because it is unrealistic to expect that people's only source of information about the population is their social circles. People interact with many others who are not part of their immediate social circles, read and watch about others in the media, etc. Yet, these results provide a clear indication that the variance of the sampled distribution affects the variance of the perceived population distribution.

Table 1.1: Results of the regression analysis for the variance of the perceived distribution in LISS panel data and GSS data on perception of racial and ethnic diversity. DV: Variance of the perceived population distribution; IV: Variance of the social circle. The results are illustrated per domain and for the panel data with characteristic fixed effects. *** - p<0.01, ** - p<0.05,* - p<0.1. Standard errors are in the parentheses.

Characteristic	V_S	Constant	N
Amount of Stress	0.10*** (0.02)	2.28*** (0.04)	1,407
Personal Income	0.12*** (0.02)	2.32*** (0.03)	1,408
Household Income	0.13*** (0.03)	2.30*** (0.02)	1,407
Wealth	0.12*** (0.03)	2.40*** (0.03)	1,404
Number of Friends	.16*** (0.02)	1.89*** (0.03)	1,408
Level of Education	0.16*** (0.03)	2.21*** (0.02)	1,410
Number of Problems	0.19*** (0.04)	2.21*** (0.02)	1,409
Number of Meetings	0.21*** (0.03)	1.78*** (0.03)	1,299
Number of Conflicts	0.13*** (0.04)	2.07*** (0.03)	1,087
Number of Dates	0.18*** (0.10)	2.51*** (0.06)	277
Pooled LISS data	0.15*** (0.01)	- -	12,516
GSS Data	0.17*** (0.02)	2.17*** (0.03)	1061

Table 1.2: Comparison between the variance of the perceived population distribution (V_P), the variance of the social circle distribution (V_S) and the variance of the real population distribution (V_R) in LISS Panel Data. 95% Confidence intervals are in the brackets.

	$P(V_P < V_R)$	$P(V_S < V_R)$	$P(V_P < V_R \mid V_S < V_R)$	$P(V_P < V_R \mid V_S > V_R)$	Difference in Prop.
Amount Stress	.87	.96	.88	.80	.08 [-.03,.20]
Personal Income	.52	.90	.53	.37	.16 [.07, .25]
Household Income	.72	.94	.73	.59	.15 [.03, .26]
Wealth	.70	.94	.71	.56	.14 [.03, .26]
Number of Friends	.67	.93	.68	.47	.21 [.10, .32]
Level of Education	.26	.87	.28	.15	.13 [.06, .18]
Number of Problems	.72	.92	.74	.54	.20 [.09, .30]
Number of Meetings	.78	.93	.79	.59	.20 [.09, .30]
Number of Conflicts	.33	.88	.35	.17	.18 [.11, .26]
Number of Dates	.20	.74	.24	.07	.17 [.08, .27]
Average	.61	.92	.64	.39	.25 [0.22, 0.28]

It is possible to provide further evidence for the hypothesis that sample variability affects perceived population variability by testing another prediction of our sampling mechanism: that the perceived population variability will be *lower* than the true variability for most respondents. For each characteristic, respondents were asked to indicate their own position on the seven-level scale. This allowed us to construct the true population distribution.[11][12]

Table 1.2 reports the proportion of respondents for which the perceived population distribution had a variance lower than the variance of the true population distribution (see column $P(V_P < V_R)$). This is higher than 50% for most characteristics. In the pooled data, the proportion of underestimation was 0.61. In summary, there is a general tendency toward underestimating the variability of the population distributions.

There is a similar pattern regarding social circle distributions. For most participants, the variance of the social circle distribution was lower than the variance of the true population distribution. This was the case for all 10 characteristics, as well as for the pooled data (see Table 1.2, column $P(V_S < V_R)$).

Our model predicts that the probability of underestimation of the variance of the real distribution depends on whether the subject's sample underestimates it. To test this prediction, we computed two conditional probabilities: the probability that a subject underestimates the variance of the real distribution when her sample underestimates it, $P(V_P < V_R \mid V_S < V_R)$, and when it overestimates it, $P(V_P < V_R \mid V_S > V_R)$. Our model predicts that the first is larger than the second. That is what we find. In the pooled data $P(V_P < V_R \mid V_S < V_R)$ across all characteristics is 0.64 whereas $P(V_P < V_R \mid V_S > V_R)$ is only 0.39. This the same pattern holds for all 10 characteristics (see Table 1.2).

Overall, these analyses of the LISS data provide evidence that information sampling likely plays a role in shaping the variability of perceived population distributions.

Analysis of GSS data

We replicated our analyses of the LISS data using another dataset: the GSS survey data. The General Social Survey is collected by the National Opinion Research Center at the University of Chicago and is based on a representative sample of US citizens (Davis et al., 2016). In the year 2000 edition of the survey, data on per-

[11] The data were collected in two waves. In the first wave, participants reported the perceived population distribution and in the second wave participants reported the social circle information. In each wave, participants indicated their position in the distribution. The results discussed in the text are based on a real distribution constructed from the responses (of participants about their position) collected in the first wave. The results are essentially the same for the distribution based on the second wave responses and thus we do not report these additional analyses here.

[12] There is a bit of irony in calling the 'true population distribution' a distribution constructed on the basis of a sample of smaller size than the true population (the population of the Netherlands). But because this sample is large, (about 1,400 people), its sample variance is very likely to be almost identical to the population variance.

ception of racial and ethnic diversity was collected. Similarly to the LISS panel, respondents were asked to report the percentage of different races and ethnicities in their communities and in the US population. The collected data contains details about major ethnic groups: Whites, Black, Hispanics, Asians, Jews and American Indians. We estimated the variance of the perceived distribution of mentioned groups in the community. We did the same for the perceived distribution of mentioned groups in the US population.[13] Linear regression of the variance of the perceived population distribution on the variance of the distribution in the community leads to results similar to what we obtained with the LISS data: The higher the variance is in the community, the higher the variance of the perceived population distribution is (see Table 1.1 for details). Under the assumption that the perceived community distribution closely corresponds to distribution sampled by the respondents, this result indicates that sample variability and perceived variability of the population are positively associated. [14]

Summary

We have provided direct (from experimental data) and indirect (from the analysis of survey data from nationally representative panels of respondents) that sampled variability affects perceived variability. Next, we review existing evidence and report new evidence about the effect of sample size on perceived variability.

1.9.2 Sample Size and Perceived Variability

Existing Evidence

Linville et al. (1989) asked Yale undergraduate students to estimate the distribution of their classmates in an introductory psychology class on 5 characteristics: likability, average number of hours per day spent studying outside of class, SAT scores, typical mood, and friendliness (Experiment 4). All characteristics had 7 levels. For each characteristic, participants were asked to indicate the percentage of their classmates that fall into each of the 7 levels. The authors elicited the perceived distributions three times: near the beginning, at the midpoint and near the end of the semester. They found that perceived variability increased with experience: the linear trend was positive and significant for both variance of the perceived distribution and the probability of differentiation.

Experiment 2 in Kareev et al. (2002) provides a more direct test of the hypothesis that perceived variability tends to increase with sample size. Participants saw

[13] The survey allowed the respondents to report percentages that did not sum to 100. When computing the variance, we normalized the reported percentages to sum to 1.

[14] For the GSS data, we are unable to provide the analysis of the relationship between the variance of the perceived population distribution, the variance in the respondent's community, and the real distribution. Respondents reported their race but they only had three options: "White", "Black" and "Other". This choice set makes it impossible to provide a comparable estimate of the real variance of the distribution of the mentioned ethnic groups among the respondents.

two populations of equal variance (this was unknown to the participants) and then they were asked to indicate which of the two was the less variable. The stimuli were the same as in their Experiment 1 (discussed earlier in this section). Participants were asked to judge which of the two populations was more variable (on a unique dimension). Unbeknown to the participants, the two populations had the same distribution. They saw a sample from each population. For one population, participants saw the whole population (28 items). We call it the 'large sample population'. For the other, they draw a random sample of 7 items. We call it the 'small sample population.' The majority of participants indicated the small sample population as the less variable. Participants also completed an incentivized task where the optimal choice was to select the less variable population (they were told that two items will be drawn from the selected population and that they would receive a bonus if they were close enough). Again, the majority of participants selected the small sample population. Overall, these results indicate that the participants perceived the small sample population as less variable than the large sample population.

These two studies provide evidence that the perceived variability of a distribution increases with sample size. Yet, in neither study sample variability is actually analyzed. For example, without the information about the actual samples observed by the participants, it is not impossible to rule out the possibility that the people perceive a large sample population as more variable even if the observed sample is not more variable (in the Kareev et al. experiment). The study by Linville et al. (1989) is subject to the same limitation. To overcome this limitation of currently available evidence, we ran a new experiment.

Experiment

Design Our design is inspired by features of the experiment in Goldstein and Rothschild (2014) and of Experiment 2 in Kareev et al. (2002). The flow of the experiment was as follows. After seeing the consent form, participants received general instructions: "Imagine we have two extremely large bags: one with RED ping pong balls and one with BLUE ping pong balls. Each ball (both red and blue) has a value between 1 and 10 written on it. During the experiment, you will observe balls first from one bag and then from another. In the end, you will have to judge which bag has the larger variety of numbers on the balls." Then, participants observed a random sample from one bag and in the following block a random sample from the other bag. The sample sizes were 5 and 50. The pairing of the color and the sample size was randomized as well as the order in which the two samples were presented. The samples were drawn from the same distribution. We used a symmetrical distribution which ranged from 2 to 9 with the following frequencies: $[0.01, 0.06, 0.17, 0.26, 0.26, 0.17, 0.06, 0.01]$. This distribution is a rescaled and discretized beta distribution with parameters $\alpha = \beta = 5$. Each participant observed a unique random sequence from the distribution. Before each sample, the participants saw a fixation cross for 450 milliseconds. Then digits appeared on the screen

in quick succession (each digit remained on the screen for 600 milliseconds).

After participants observed the samples from the two bags, they answered three questions pertaining the perceived variabilities of the two bags.

- ***Q1***: This question was incentivized. Participants were told: "Suppose you select two balls from one of the two bags. Let us call A and B the numbers on the balls. Let D be the difference between these two numbers. You will get a bonus of D points. That is, the larger the difference between the two numbers, the higher your bonus (the bonus cannot be negative)." At the end of the experiment, two balls were randomly drawn from the chosen bag and participants were paid a bonus proportional to D. The goal was thus to select the bag with the higher variability.

- ***Q2***: Participants were presented with a continuous slider where they indicated which bag had the larger "variety of numbers on the balls". The minimal value of the slider was -100 (e.g., 'The Red bag has more variety'). The maximal value was 100 (e.g.,'The Blue bag has more variety') and had a midpoint at 0 (e.g.,'The Red and Blue bags have the same variety'). (The colors at the end of the scales were randomized and the numeric values were not shown to the participants).

- ***Q3***: Participants were asked to imagine they would pick two balls from each of the two bags. Then they were asked to indicate the bag for which they predicted the two numbers to be closer to each other.

We recruited 303 participants using Amazon Mechanical Turk. Participants received fixed payment for their time and a bonus based on their responses to question Q1.[15]

Predictions *Manipulation check:* We anticipated that for most participants the sample variability of the large sample bag (V_L^c) would be larger than the sample variability of the small sample bag (V_S^c): $P(V_L^c > V_S^c) > .5$. *Prediction about perceived variability:* Most participants will select the large sample bag as the more variable bag. *Prediction about the effect of sample variability:* The proportion of participants choosing the large sample bag will be higher when the large sample bag has the higher variability than when it has the lower variability.

Results The results are consistent with our prediction. We report our analyses by using the corrected sample variance as the estimator of sample variability. Similar results hold with the other estimators discussed above.

Manipulation check: For 57% of the participants, the corrected sample variance of the large sample bag V_L^c was larger than the corrected sample variability of the small sample bag V_S^c: $P(V_L^c > V_S^c) = .57, 95\%CI = [0.51, 0.63]$.

[15]Experimental data will be made available on Open Science Framework upon publication of the paper.

Perceived variability: Most participants perceived the large sample bag as more variable than the small sample bag. For Q1, 61% of the participants chose the bag of which they observed a larger sample. This proportion is significantly above 50% ($95\%CI = [0.55, 0.67]$). For Q2, 70% of the participants selected a response on the scale that indicated that the large sample bag had "more variety" ($95\%CI = [0.65, 0.75]$). The mean response was 33.33 ($95\%CI = [26.5, 40.2]$). This is significantly higher than the mid-point of 0. For Q3, 53% of the participants indicated that balls from the bag of which they observed a smaller sample were closer to each other. This proportion is only marginally significantly different from 50% ($95\%CI = [0.48, 0.59]$, $p = 0.13$).

Effect of sample variability: We computed the proportion of participants who chose the large sample bag as the more variable when its sample variance was larger. For Q1 it is .68 ($95\%CI = [0.61, 0.75], n = 173$). The corresponding proportion conditional on the larger bag having the lower sample variance is .52 ($95\%CI = [0.43, 0.6], n = 130$). The difference in proportions is significantly higher than 0: $d = 0.17, 95\%CI = [0.05, 0.28]$. Similar results hold for Q2 and Q3 (see Table 1.3).

Table 1.3: Proportion of participants who indicated the large sample bag as the more variable. 95% Confidence intervals are in the brackets.

Question	All Observations	Large Sample Bag has Larger Sample Variance : $V_L > V_S$	Large Sample Bag has Smaller Sample Variance: $V_L < V_S$	Difference in Proportions
Q1	.61	.68	.52	.17
	[.55, .67]	[.61, .75]	[.43 .60]	[.05, .28]
Q2	.70	.79	.58	.21
	[.65, .75]	[.72, .85]	[.49, .67]	[.10, .32]
Q3	.53	.64	.40	.24
	[.48, .59]	[.56, .71]	[.32, .49]	[.12, .35]
# part.	303	173	130	–

Summary

Most participants perceived the large sample option as more variable than the small sample option even though the samples were generated from the same underlying distribution. Sample size had a positive effect on sample variability and the difference in sample variabilities had a positive effect on the difference in perceived variabilities. The tendency to perceive the large sample option as the more variable is thus at least partly explained by the difference in sample variabilities.

1.9.3 Discussion

The empirical evidence from experiments and field data support our assumption that perceived variability depends on sample variability. Although the evidence discussed here does not concern the size of the samples collected about distinct social groups, it provides strong support for a key building block of our model. A more direct test of our theory would require the collection of data about the distribution of experiences with the in-groups and the out-groups as well as about perceived variability of these groups. This is an interesting avenue for future research.

1.10 Theoretical Implications

Most sampling explanations of judgment biases assume that people are 'naive intuitive statisticians:' they would process sampled information correctly, but would not be able to adjust their beliefs to reflect the sampling constraints from their environment (Fiedler & Juslin, 2006b; Fiedler, 2012). The argument advanced by Linville et al. (1989) relies on this naiveté assumption because it works only for statistically biased estimators and breaks down if the estimator of group variability is unbiased, as explained in the Introduction. A non-naive decision maker would correct for the bias induced by sample size and would instead use the corrected sample variance. Here we return to the claim that our analyses show that an in-group heterogeneity effect can emerge even when the naiveté assumption is relaxed and thus can be produced by the structure of the environment.

1.10.1 Rational Information Processing and the In-Group Heterogeneity Effect

We want to claim that the tendency to perceive the in-group as more variable than the out-group is not necessarily a consequence of biased information processing but that instead that the structure of the environment is sufficient to produce it. By 'structure of the environment', we refer the fact that the samples of information obtained about the in-group tend to be larger than the samples of observations obtained about the out-group.

Making this claim requires that we define what we mean by *'rational' (or unbiased) information processing*. Following the precepts of a 'rational analysis' of cognition delineated by Anderson (1991), we assume that the agent has a goal, some information about the structure of the environment, and processes information in a way that helps her achieve this goal. In other words, we assume that the agent is solving an optimization problem and that the optimal solution to this problem is the outcome of 'rational information processing.' We assume that the structure of the environment is that of the baseline model: The agent obtains two observations from each group before the first period, and then in each period, she obtains an observation from either the in-group (with probability r) or the out-group (with probability $1-r$). We need to make additional assumptions about what the agent knows about the structure of the environment and her goal.

Consider a scenario, where the agent knows the distributions of the focal feature in the two groups are Normal distributions but the means and variances are unknown. Her goal is to obtain, for each group, an estimator with uniformly minimum variance (UMVUE, sometimes referred to as the 'best unbiased estimator') that is unbiased: $E[V_g] = \sigma_g^2$. In other words, the agent tries to find the best solution to a constrained optimization problem. The (unique) estimator that is the best solution to this optimization problem is the corrected sample variance (Casella & Berger, 2002, Ch. 7, p. 346).

In the second scenario, suppose the agent knows the distributions of the focal features are Normal, with known means, but unknown variances σ_{in}^2 and σ_{out}^2. The agent knows the variances are random draws from a Uniform $U(0,1)$ distribution. The agent's goal is to obtain an estimator of the variance of the feature distribution that has the minimum mean square error (MMSE). The estimator that is the best solution to this optimization problem (the 'Bayesian Estimator') is the mean of the posterior $V_{t,g}^{Bayes}$.

In the 'Simple Model' Section, we showed that an in-group heterogeneity effect emerges with these estimators. Rational information processing thus implies the emergence of the in-group heterogeneity effect. In other words, the in-group heterogeneity effect does not have to be the consequence of biased information processing. In environments that lead agents to obtain large samples about in-groups than about out-groups, the in-group heterogeneity effect is an *inherent* consequence of the structure of the environment.

1.10.2 On the Possibility to Correct for the In-Group Heterogeneity Effect

Does it mean that it is not possible to prevent the in-group heterogeneity effect from occurring? After all, a rational agent should know that she will be more likely to perceive the in-group as more variable than the out-group. And she should be able to do something about it. In the context of the two scenarios considered in the previous subsection, there is no rational basis for applying a correction to the variability estimates. If the agent applied a correction that would eliminate the in-group

heterogeneity effect, the resulting estimator would no longer be a solution to the optimization problem - it would not be the result of rational information processing any longer, *in the context of the optimization problems as defined above*. For example, the agent could add an additional corrective term to the corrected sample variance that depends on sample size (the lower the sample size, the higher the corrective term). This might reduce the asymmetry: $P(V_{in} > V_{out})$ would be closer to $P(V_{t,in} < V_{t,out})$. But this would make the estimator biased. The resulting estimator would no longer be a solution of the optimization problem formulated above because the problem formulation required the estimator to be unbiased. Thus, it would not be the result of rational information processing (with respect to the objective to obtain an unbiased estimator with uniformly minimum variance).

It is important to note that an estimator that does not produce an in-group heterogeneity effect could be the solution to another optimization problem. Thus, it could be rational with respect to this alternative formulation. For example, consider a setting where no prior is available and the goal of the agent is to minimize $\Delta_t = P(V_{t,in} > V_{t,out}) - P(V_{t,in} < V_{t,out})$ (relaxing the constraint that the estimator be unbiased). A group variability estimator that would be a solution to this problem would be the corrected sample variance based on the two initial observations (periods -1 and 0):

$$V_{t,g} = V^c_{0,g} = \frac{1}{2-1}\left((x_{-1,g} - \bar{x}_{0,g})^2 + (x_{0,g} - \bar{x}_{0,g})^2\right) = (x_{-1,g} - \bar{x}_{0,g})^2 + (x_{0,g} - \bar{x}_{0,g})^2, \tag{1.17}$$

where $\bar{x}_{0,g} = 1/2(x_{-1,g} + x_{0,g})$. This estimator is produced by rational information processing given the optimization problem we defined (i.e., the goal of the agent and the constraints on the set of possible estimators). But this estimator does not seem quite right: it only uses the first two observations. It makes partial use of the data in the sample, which implies it cannot be of uniform minimum variance. This implies that although it is possible to construct an estimator that is rational *by some definition of rationality* and does not produce an in-group heterogeneity effect, this estimator does not seem appropriate because it is the solution to an optimization problem that does not seem right. The issue is that this problem relaxes the constraints that the estimator is unbiased and of minimum variance. But these are generally seen as two desirable constraints of point estimators (in settings where a prior is not available).

In summary, our claim that rational information processing leads to the emergence of the in-group heterogeneity effect is contingent on the (possibly implicitly) assumed goal of the agent. If we assume that the distributions are normal, the agent wants to have an unbiased estimate of group variability and to minimize the variance of the estimator, the in-group heterogeneity effect will emerge as an outcome of rational information processing. It might not emerge if the goal of the agent is to minimize the in-group heterogeneity effect (with no other constraint).

1.10.3 On the Naive Intuitive Statistician

An important interpretation of the fact that the in-group heterogeneity effect can emerge as a consequence of rational information processing is that it can emerge even when relaxing the naiveté assumption usually made by sampling explanations for judgment biases. Importantly, this does not imply anything about the extent to which people are 'naive.' A large amount of research has shown that people are oblivious to sampling biases and that even when they want to correct for those, they do not succeed. In other words, they would lack the 'metacognitive abilities' to correct for these sampling constraints (Fiedler, 2012).

We do not want to claim that people are not 'naive' or not subject to metacognitive myopia. In fact, our experiment shows that people did not apply a proper systematic correction to their variability estimates. If they had done so, the mean response to the second question would have been 0, but instead it was 33.3 ($95\%CI = [26.5, 40.2]$). In other words, it was biased toward the alternative about which a larger sample was obtained.

Our result about rational information processing has distinctive normative implications, however. It implies that making people aware of the sampling asymmetry and giving them the tools to overcome their metacognitive myopia might not be enough to eliminate the in-group heterogeneity effect. The fact that the structure of the environment is sufficient to produce the in-group heterogeneity effect implies that attempts at eliminating it require not only helping people process available information better but also changing people's sampling environment.

1.10.4 Relation to Existing Rational Analyses

At an abstract level, the mechanism underlying our results is similar to the mechanism at the core of the rational analysis of learning-by-doing by Le Mens and Denrell (2011). In that paper, the authors studied an asymmetric bandit problem with two alternatives that differed in terms of availability of outcome information. In a bandit problem, the goal of the decision maker is to maximize her expected cumulative payoff obtained over a set of periods by selecting one of the two options in each period. The task is difficult because parameters of the payoff distributions are unknown to the decision maker. When the decision maker selects an alternative, she obtains some payoff that accrues toward her earnings and she also obtains information that she can use to update her beliefs about the payoff distribution of the selected alternatives. Le Mens and Denrell studied a version of the bandit problem where, for one alternative (S), feedback information was obtained only if it was selected, whereas feedback information was obtained in every period for the other alternative (I). They showed that most rational decision makers would come to believe I to be superior to S. In this setting, a rational decision maker had full knowledge of the structure of the environment, correct prior, processed information using Bayes theorem and used the choice policy that maximized the expected payoff.

Just as in our analyses of rational information processing, rational decision makers in the asymmetric bandit problem had unbiased beliefs. Yet, most of them came to evaluate I more positively than S. Even if a rational decision maker were aware of this tendency, she would have no rational basis to change her belief, nor her sampling behavior. This is because changing her sampling behavior would conflict with her payoff maximizing goal. The intuition for this phenomenon is that the Bayesian estimator for alternative S has a skewed distribution (and is unbiased). This is similar to what happens in our setting with the variability estimators. Unbiased variability estimators are skewed. Because the strength of the skew changes with sample size, and the size of the sample about the in-group tends to be larger than the size of the sample about the out-group this implies that the estimator of the difference in variabilities ($V_{in} - V_{out}$) will also be skewed (but unbiased).

The role of the skew was also noted by Kareev (1995, 2000) in analyses of intuitive estimates of correlation based on small samples: the distribution of sample correlations based on small samples is skewed in such a way that correlations are overestimated most of the time when the sample size is about 7 (under certain assumption). As noted by Le Mens and Denrell (2011), a rational agent aware of this fact would not want to use a potentially corrected estimator because the correction would induce a bias.

More generally, we conjecture that sampling explanations of systematic judgment asymmetries become possible in settings where the environment produces information samples that lead to skewed distributions of beliefs. Uncovering new phenomena and corresponding environmental influences is an exciting avenue for future research.

1.11 Conclusion

People frequently obtain larger samples of information about in-groups than about out-groups. Because estimators of variability tend to be right-skewed when samples are not very large, people will tend to perceive in-groups as more variable than out-groups. In this paper, we showed that this in-group heterogeneity effect emerges under a wide range of assumptions about how people process information. In particular, it emerges even when people process information rationally. This implies that sampling constraints imposed by the environment are sufficient to imply the emergence of the in-group heterogeneity effect. This sampling explanation complements existing explanations that focused of how information about in-group and out-group members is processed.

Chapter 2

FEATURE INFERENCE WITH UNCERTAIN CATEGORIZATION: RE-ASSESSING ANDERSON'S RATIONAL MODEL

Joint with Gaël Le Mens
Forthcoming in ***Psychonomic Bulletin and Review***

2.1 Introduction

According to J. Anderson, 'The basic goal of categorization is to predict the probability of various inexperienced features of objects' (Anderson, 1991). At the same time, humans often find themselves in situations where the categories of the objects they perceive are uncertain. How do people make predictions about unobserved features of an object when the category of that object is uncertain?

A highly influential answer to this question is J. Anderson's 'rational model' (Anderson, 1991). Consider a setting where an individual observes a feature of an object and makes a prediction about an unobserved feature of that object. The values of the two features are denoted by X (first feature) and Y (second feature). It is assumed that the individual has organized her knowledge of the domain in a set of categories $\mathcal{C}$.[1] According to Anderson's model (AM), the probability that the value of the second feature is y when the individual knows that the value of the

[1] When we refer to 'Anderson's model' we only refer to the feature inference component of the original model. The original model had additional components to allow it to learn categories. This way of referring to Anderson's model is similar to Murphy and Ross (2010a).

first feature is x is given by

$$P(y \mid x) = \sum_{c \in C} P(c \mid x) P(y \mid c), \tag{2.1}$$

where $P(c \mid x)$ is the subjective probability that the object comes from category c given the observed feature value x and $P(y \mid c)$ is probability that the second feature has value y given that the object belongs to category c. An important qualitative prediction of this model is that people take into account all the candidate categories when making an inference about the unobserved feature Y on the basis of the value of the observed feature $X = x$.

A large amount of empirical work has focused on testing this prediction. Existing findings are mixed. Some experimental evidence suggests that participants' inferences are the same as those implied by a model that relies only on the most likely category given the observed feature (the 'target category') (Chen, Ross, & Murphy, 2014a, 2014b; Malt, Ross, & Murphy, 1995; Murphy & Ross, 1994, 2010a; Murphy, Chen, & Ross, 2012; Ross & Murphy, 1996; Verde, Murphy, & Ross, 2005). Other experiments suggest that participants rely on more than just the target category (Chen et al., 2014a; Hayes & Chen, 2008; Hayes & Newell, 2009; Murphy & Ross, 2010a; Newell, Paton, Hayes, & Griffiths, 2010; Verde et al., 2005). Finally, still other experiments suggest that participants do not pay attention to categories at all but instead are sensitive to the overall feature correlation (O. Griffiths, Hayes, Newell, & Papadopoulos, 2011; O. Griffiths et al., 2012; Hayes, Ruthven, & Newell, 2007; Papadopoulos, Hayes, & Newell, 2011). Several recent papers have attempted to uncover the conditions under which people are more likely to rely on multiple categories or just the target categories. For example, Murphy and Ross (2010b) found that participants were more likely to use multiple categories when the most likely category gives an ambiguous inference, and less likely to do so when the most likely category gives an unambiguous inference. Chen et al. (2014a) found that participants' inferences were likely to be influenced by multiple categories when the inference was implicit whereas they were likely to be influenced by just the target category when the inference was explicit. O. Griffiths et al. (2012) found that participants' inferences were more likely to be influenced by a single category when participants had been trained to classify stimuli before the feature induction task.

Despite the diversity of findings, the studies that analyzed the performance of Anderson's model converge in showing that it provides a poor fit to experimental data. Central to this model is an assumption about the structure of the environment: it assumes that the within-category feature correlations are equal to 0 (this is the '*conditional independence*' assumption). We believe this model can be seen as a 'rational model' only to the extent that this assumption is consistent with the structure of the actual task environment. We reviewed all prior experiments on feature inference with uncertain categorization (reported in the papers cited above) to check whether the task environments of these experiments were characterized by conditional independence. We found that it is the case in *none* of the previously

published experiments.[2]

The poor performance of Anderson's model in an environment without conditional independence suggests that people do not make this assumption in such environments (a point made by Murphy & Ross, 2010a). Yet, currently available evidence provides little information about how this model would fit participants' inferences in a setting where conditional independence is satisfied. How well would Anderson's model (AM) predict participants' inferences in a task environment consistent with the conditional independence assumption?

At first sight this question might seem moot. After all, Murphy and Ross (2010a) noted that there are many environments in which this assumption is not satisfied. For example, they argued that within category feature correlation can result from large category difference. One example is sexual dimorphism in animals (Murphy & Ross, 2010a, p. 14). Male deer are larger and have different coloration than females. Therefore, these features are correlated within the category 'deer'. Similar feature correlations are present in consumer goods categories like books or computers. There is also evidence that people are aware of some within category correlations (Malt & Smith, 1984).

However, even if there are possibly few naturally occurring environments that satisfy conditional independence, it is important to assess the performance of the Anderson's model in such settings. This is because there currently does not exist a rational model for environments where the conditional independence assumption does not hold. If Anderson's model performs well under conditional independence – when it can be seen as a 'rational model' – this will suggest that an extension of this model to settings without conditional independence needs to be developed. Such a model is likely to perform well.

We analyzed the performance of Anderson's model in a task environment characterized by conditional independence, consistent with this key assumption of the model. In 5 experiments, we found that the model performed better than other competing models. This finding is important because it suggests that people's inferences can be influenced by several categories when making inferences under uncertain categorization. Although there already exists some evidence that this can be the case (e.g., Chen et al., 2014a; O. Griffiths et al., 2012; Murphy & Ross, 2010b), we explain below that such evidence is based on a design that does not allow the parsing out between two possible interpretations of the data: that participants ignore categories altogether or that categories influence inferences in a fashion close to what would be predicted by application of Bayes' theorem. The

[2]In Hayes and Newell (2009), the authors write that 'all the experimental categories were designed so that their component feature dimensions were statistically independent within and between categories' (p.733). Computations based on the statistical structure reported in their Table 1 show that this claim is not exact. Consider the Terragaxis category in the 'Divergent' condition. Here we have P(Rash & Headache) = 3/8, P(Rash) = 4/8 and P(Headache) = 7/8. If Rash and Headache were independent within category, we would have P(Rash & Headache) = P(Rash) P(Headache). This is clearly not the case. Therefore, the statistical structure of the experiments in this paper is not characterized by conditional independence. (This critique does not invalidate in any ways the results reported in that paper.)

results reported in this paper suggest the later interpretation.

In the following, we describe the existing experimental paradigm that has been used by most of the literature on feature inference under uncertain categorization. We explain how the fact that it relies on discrete-valued features makes it of limited usefulness to assess the performance of Anderson's model. Then we introduce our adaptation of Anderson's model to continuous environments and describe competing models. Subsequently, we report the performance of Anderson's model in 4 experiments based on a novel paradigm with continuous features and one experiment based on the existing paradigm with discrete features. Finally, we discuss how our findings relate to prior research.

2.2 Existing Paradigm - Discrete Features

In the experimental paradigm used in the vast majority of experiments that focused on feature prediction with uncertain categorization, participants are shown a set of items of various shapes and colors divided into small number of categories, typically 4 (Murphy & Ross, 1994). Then they are told that the experimenter has a drawing of a particular shape and were asked to predict its likely color (or similar questions about the probability of an unobserved feature given an observed feature). An important characteristic of this paradigm is that the categories are shown graphically to the participants. The idea was to avoid complications related to memory and category learning by participants.

Suppose the two features are X and Y and there are 4 categories. Participants are asked to estimate $P(y \mid x)$, the proportion of items with $Y = y$ out of items with $X = x$. There is some evidence that participants' predictions are the same as those implied by a model that focuses on just the 'target' category, that is, the most likely category given the observed feature (Murphy & Ross, 1994). There is also some evidence that participants sometimes make predictions that are the same as those implied by a model that takes into account multiple categories (Murphy & Ross, 2010a). Still, other experiments have found evidence that participants do not pay attention to categories at all but instead are sensitive to the overall feature correlation (Hayes et al., 2007; Papadopoulos et al., 2011; O. Griffiths et al., 2012).

A limitation of this paradigm pertains to the fact that the features are discrete-valued. This implies that the predictions of a model that ignores categories altogether or makes optimal use of the categories are exactly the same. This is a consequence of the law of total probability. In this case, we have

$$P(y \mid x) = \sum_{c=1}^{4} P(c \mid x) P(y \mid cx), \tag{2.2}$$

where $P(c \mid x)$ is the proportion of items belong to c out of all the items such that $X = x$, and $P(y \mid cx)$ is the proportion of items with $Y = y$ out of the items that both are in c and have $X = x$.

In settings where there is conditional independence, we have $P(y \mid cx) = P(y \mid c)$ and thus the above equation can be rewritten as:

$$P(y \mid x) = \sum_{c=1}^{4} P(c \mid x) P(y \mid c). \tag{2.3}$$

In order to estimate $P(y \mid x)$, a participant that would ignore the categories would consider all objects with $X = x$ and would respond with the proportion of objects with y among all objects with x. A participant that would consider all 4 categories would compute the proportion of items with y among the items with x *in each category* and then would compute the weighted average by multiplying each of these numbers by her estimates of $P(c \mid x)$. The responses given by the two participants would be *exactly the same*. It is therefore difficult to assess whether the participants use multiple categories (but see Murphy and Ross (2010a) for an attempt to do so using post-prediction questions). When features are continuous, however, the predictions of these two strategies differ.

Below, we describe a version of Anderson's model adapted to a continuous environment and report 4 experiments designed to test this model. We return to the discrete environment setup in Experiment 5 and the General Discussion section.

2.3 Rational Feature Inferences in a Continuous Environment

2.3.1 Representing Mental Categories

We depart from the prior literature on feature inference with uncertain categorization by focusing on a setting with continuously valued (as opposed to discrete) features. Following recent work, we model mental categories using probability distribution functions (*pdfs*) on the feature space (Ashby & Alfonso-Reese, 1995; Sanborn, Griffiths, & Shiffrin, 2010). Let $c \in \mathcal{C}$ be a category. We denote by $f(x,y \mid c)$ the value of the associated *pdf* at position (x,y) in the feature space, where x denotes the value of the first feature and y denotes the value of the second feature. This *pdf* denotes the prior belief of the individual over positions given that she knows that an object is from category c.

For simplicity, in what follows we assume there are two relevant categories ($\mathcal{C} = \{1,2\}$) each represented by bi-variate normal distributions (Ashby & Alfonso-Reese, 1995):

$$\begin{pmatrix} X_c \\ Y_c \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_{xc} \\ \mu_{yc} \end{pmatrix}; \begin{pmatrix} \sigma_{xc}^2 & 0 \\ 0 & \sigma_{yc}^2 \end{pmatrix} \right), \tag{2.4}$$

where μ_{xc} and μ_{yc} are the category means for the two features, and σ_{xc} and σ_{yc} are the standard deviations. Consistent with the conditional independence assumption, the within-category feature correlation is zero. See Figure 2.1 for an example.

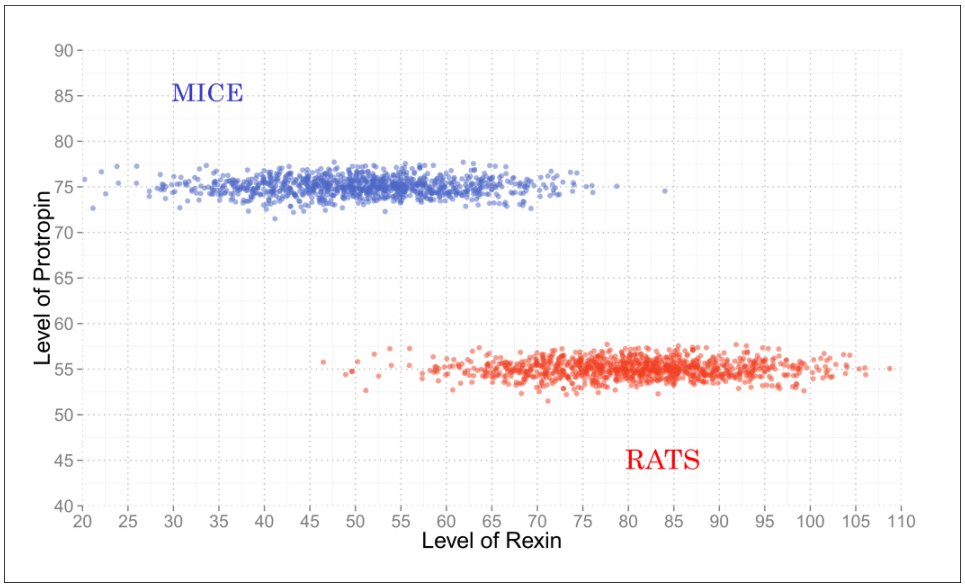


Figure 2.1: Categories used in the 4 experiments. Participants were shown the level of 'Rexin' (x-axis) and were asked to predict the level of 'Protropin' (y-axis). The categories are 'rat' (R) and 'mouse' M). $\mu_{xR} = 80, \mu_{yR} = 55, \mu_{xM} = 50, \mu_{yM} = 75, \sigma_{xR} = \sigma_{xM} = 10, \sigma_{yR} = \sigma_{yM} = 1$.

2.3.2 Anderson's Rational Model (AM)

By adapting equation 2.1 to this continuous setting, we express the posterior on the second feature given the value of the first feature:

$$f(y \mid x) = \sum_{c \in C} P(c \mid x) f(y \mid c), \tag{2.5}$$

where $P(c \mid x)$ is the subjective probability that the object comes from category c given that the first feature is observed to have value x and $f(y \mid c)$ is the marginal distribution of the second feature, conditional on the fact that the object is a c.

Anderson's model assumes that the subjective probabilities of the candidate category are given by Bayes' theorem:

$$P(c \mid x) = \frac{P(c) f(x \mid c)}{f(x)} = \frac{P(c) \int_v f(x, v \mid c) dv}{\sum_{c \in C} P(c) \int_v f(x, v \mid c) dv}, \tag{2.6}$$

where $P(c)$ is the prior on the category.

In the special case with two categories and normally distributed category *pdfs*, we have:

$$f(y \mid x) = P(c_1 \mid x) f_{\mu_{y1}, \sigma_{y1}}(y) + P(c_2 \mid x) f_{\mu_{y2}, \sigma_{y2}}(y), \tag{2.7}$$

where f_{μ_y,σ_y} denotes the density of a normal distribution with mean μ_y and standard deviation σ_y, $P(c_2 \mid x) = 1 - P(c_1 \mid x)$, and

$$P(c_1 \mid x) = \frac{1}{1 + e^{ax^2 - bx + c}}, \tag{2.8}$$

with

$$\begin{aligned}
a &= \frac{\sigma_{x2}^2 - \sigma_{x1}^2}{2\sigma_{x2}^2\sigma_{x1}^2}, \\
b &= \frac{\sigma_{x2}^2\mu_{x1} - \sigma_{x1}^2\mu_{x2}}{\sigma_{x2}^2\sigma_{x1}^2}, \\
c &= \frac{\sigma_{x2}^2\mu_{x1}^2 - \sigma_{x1}^2\mu_{x2}^2}{2\sigma_{x2}^2\sigma_{x1}^2} + \log\frac{\sigma_{x2}}{\sigma_{x1}} + \log\frac{P(c_2)}{P(c_1)}.
\end{aligned}$$

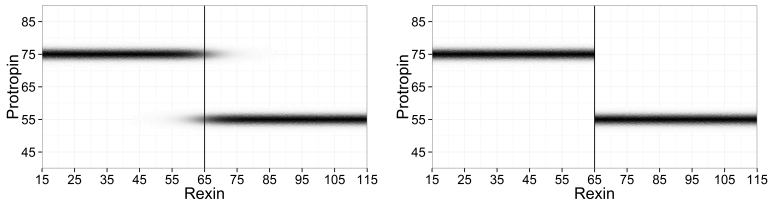
We assume that the priors on the two categories, $P(c_1)$ and $P(c_2)$, are both equal to 0.5.

2.3.3 Competing Models

Prior literature suggests that people frequently focus on the most likely category and that they sometimes ignore categories altogether but are sensitive to the overall feature correlation. We describe 'translations' of these perspectives to the continuous environment.

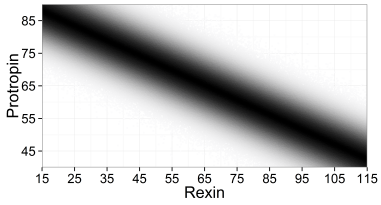
Single Category - Independent Features (SCI) We refer to the most likely category given the observed feature (x) as the 'target' category (this is category 1 if $P(c_1 \mid x) > .5$, as given by eq. 3.9). The posterior has the same structure as in Anderson's model, but with all the weight on the target category (c^*). In this case, $f(y \mid x) = f_{c^*}(y \mid x)$, where $f_{c^*} = f_{\mu_{y1},\sigma_{y1}}$ if the target category is category 1, and $f_{c^*} = f_{\mu_{y2},\sigma_{y2}}$ otherwise. The 'switch' is situated where x is such that $P(c_1 \mid x) = .5$. In the rest of the paper we refer to this value as the 'boundary'.

Linear Model (LM) Prior literature considered the 'feature conjunction' approach as a model that is sensitive to the overall statistical association between the two features across objects, independently of categorical boundaries. This model simply computes the empirical probability of the unobserved feature given the observed feature based on all the data, ignoring categorical boundaries. A direct analogue in the continuous setting does not exist because the agent might have to infer Y conditional on an x value to which she has never been exposed. This observation implies that a model that 'regularizes' the available observations is in order. This could be a parametric model or a non-parametric exemplar model that weights prior observations based on their similarity to the stimulus (Ashby & Alfonso-Reese, 1995; Nosofsky, 1986). For the sake of simplicity, we analyze a



(a) AM: Anderson's Model.

(b) SCI: Single Category Independent Features



(c) LM: Linear Model

Figure 2.2: Posteriors ($f(y \mid x)$) on the second feature (Rexin) given the value of the first feature (Protropin) of the competing models. The darker areas correspond to higher values of the posterior and lighter areas correspond to lower values. The vertical line at $x = 65$ represents the 'boundary', i.e., the value of x for which the two categories are equally likely $P(\text{'rat'} \mid x = 65) = P(\text{'mouse'} \mid x = 65)$.

linear model. This is the simplest model that takes into account the overall feature correlation:

$$f(y \mid x) = f_{a_0+a_1x,\sigma_l}(y), \tag{2.9}$$

where $f_{a_0+a_1x,\sigma_l}(y)$ denotes a normal *pdf* with mean $a_0 + a_1x$ and standard deviation σ_l. The parameters are the coefficients of the best fitting linear model based on the observed samples from the two categories.

2.3.4 Decision Rule

The outputs of all three models, as described above, are posterior distributions: subjective probability distributions over the value of the second feature (Y) given the observed feature (X). To make empirical predictions about human inferences, we need to specify how this posterior distribution translates into responses. In analyses of our experimental results, we will assume that the response is a random draw from the posterior distribution – this is a 'probability matching' decision rule. Other decision rules are theoretically possible. They would lead to different model predictions. We return to this issue in the General Discussion section of the paper.

2.4 Experiment 1

Participants faced a feature inference task that closely matches the setting of the previous section. Following standard practice in the study of feature inference with uncertain categorization, we used a 'decision-only' paradigm: participants were provided with a graphical depiction of the categories which remained visible when they made inferences about the second feature on the basis of the value of the first feature. We adopted this design to avoid issues related to memory.

2.4.1 Design

Our experiment used artificial categories to avoid the influence of domain-specific prior knowledge. We asked the participants to assume they were biochemists who studied the levels of two hormones in blood samples coming from two categories of animals (e.g., Kemp, Shafto, & Tenenbaum, 2012)). The hormones were called 'Rexin' and 'Protropin' and the two categories of animals were 'Mouse' and 'Rat.' We provided the participants with visual representations of the categories in the form of scatter plots of exemplars of the two categories (see Figure 2.1). In addition, participants went through a learning procedure designed to familiarize themselves with the position of the categories in feature space (see Supplementary Material). In the judgment stage, participants were asked to infer, without feedback, the likely level of Protropin, based on the level of Rexin, for 48 blood samples which didn't indicate the animal they came from (the category was thus uncertain). The question was 'What is the likely level of Protropin in this blood sample?' Participants answered using a slider scale with minimal value 40, maximal value 90, and increments of 1 unit.

30 participants recruited via Amazon Mechanical Turk completed the experiment for a flat participation fee.[3]

2.4.2 Model Predictions

Figure 3.5 depicts the posterior distributions, $f(y \mid x)$, implied by the three competing models. The posterior for AM is based on equation 2.7 and the Bayesian category weights given by equation 3.9. The posterior for SCI is based on equation 2.7 and the all-or-nothing category weights. The parameters are the coefficients used to generate the categories (see the legend of Figure 2.1). The posterior for LM is based on equation 2.9. The parameters are the coefficients of the best fitting linear model based on the all the dots depicted on Figure 2.1 (irrespective of their categories). With the stimuli used in the experiment, we have $a_0 = 94.8, a_1 = -0.45$ and $\sigma_l = 5.7$.

The crucial difference between the predictions of AM and SCI lies in the region around the x value at which both categories are equally likely (Rexin level of 65).

[3]The stimuli and data from the 5 experiments are available on Open Science Framework: www.osf.io/wps39.

Consider Rexin level of 60. According to SCI only high levels of Protropin (close to 75) are likely (the ones corresponding to the "Mouse" category). According to AM, however, both high (close to 75) and low levels (close 55) of Protropin are likely.

2.4.3 Results

Parameter-Free Model Comparison Here we assume that the model parameters for AM and SCI are the coefficients used to generate the categories and that the parameters for LM are those of the best fitting regression line, just as in Figure 3.5. We computed the log-likelihood fit of each model on a participant-by-participant basis.[4] Anderson's model (AM) is the best fitting model for the majority of participants (74% of them, see table 3.1). Figure 3.6 shows the inferences of all participants as well as the log-likelihood of the three models for each participant.

[4]In all analyses we removed two judgments where the given Rexin level was $x = 65$. At this value, the two categories are equally likely and the SCI does not make an explicit prediction.

Table 2.1: Percentage of participants whose feature predictions were best fit by each of the candidate models.

	Exp. 1		**Exp. 2**		**Exp. 3**		**Exp. 4**	
Model	True	Est.	True	Est.	True	Est.	True	Est.
AM: Anderson	22(74%)	18(60%)	26(87%)	26(87%)	19(63%)	17(57%)	23(74%)	21(68%)
SCI: Single Cat. Indep. Features	4(13%)	12(40%)	1(3%)	3(10%)	0	9(30%)	3(10%)	8(26%)
LM: Linear	4(13%)	0	3(10%)	1(3%)	11(37%)	4(13%)	5(16%)	2(6%)
# participants	30		30		30		31	

Table 2.2: Estimated model parameters. Parameters were estimated separately for each participant. The values are the mean estimated parameters across participants for whom that model is the best. AM: Anderson's rational model, SCI: the single category independent feature model; LM: linear model.

	Parameter	$\mu_{x,R}$	$\mu_{y,R}$	$\mu_{x,M}$	$\mu_{y,M}$	σ_x	σ_y	a_0	a_1	σ_l
	True Value	**80**	**55**	**50**	**75**	**10**	**1**	**94.8**	**-0.45**	**5.7**
Experiment	**Model**									
1	*AM*	81.4	55.2	44.9	74.9	11.5	1.0			
	SCI	82.0	55.3	45.9	74.7	12.9	1.0			
	LM							88.3	-0.4	6.6
2	*AM*	75.7	55.4	44.9	75.3	10.7	0.9			
	SCI	81.6	55.1	51.9	75.2	15.5	0.4			
	LM							24.1	0.6	3.9
3	*AM*	77.9	54.9	43.2	74.9	11.6	1.0			
	SCI	79.3	55.7	49.9	73.8	11.4	2.9			
	LM							70.6	-0.1	6.0
4	*AM*	81.7	55	48.2	75.2	17.5	0.9			
	SCI	82.7	55	44.9	74.9	24.9	0.7			
	LM							68.8	0.1	4.9

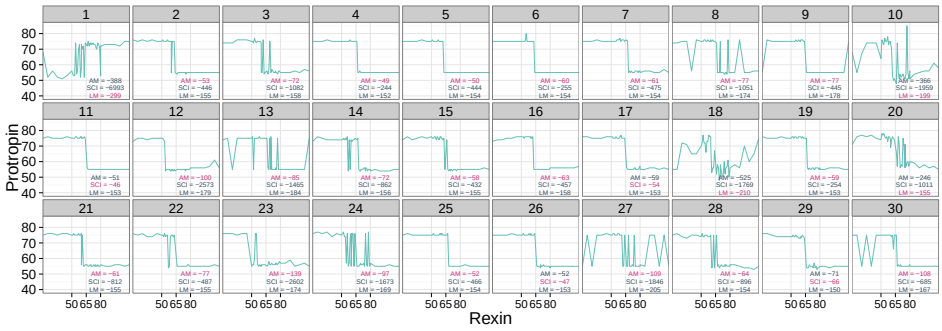


Figure 2.3: Inferences of the participants of Experiment 1. The log-likelihoods of each parameter-free model is shown on each participant's graph. AM: Anderson's rational model, SCI: single category independent feature model; LM: linear model. The red font indicates the best fitting model.

Comparison of Models with Parameters Estimated Participant-by-Participant The comparison of the parameter-free models implicitly assumes that the participants perceived the categories accurately (i.e. the parameters of their category *pdfs* were exact). This might not have been the case, however. For example, participants might have misjudged the position of the point where categories are equally likely ($x = 65$). Inspection of Figure 3.6 reveals that the perceived position of this 'boundary' is essential to the performance of the single category model (SCI). A slight error leads to a strong penalty in terms of log-likelihood that might not translate to the fact that a participant used multiple categories. For example, participant #4 made predictions that are clearly indicative of a focus on just one category since the predictions correspond to the median y-level for 'Mouse' (the category on the left) when x is low and to the median y-level for rats (the category on the right) when x is high. But the participant switched between categories not exactly at the 'boundary' of $x = 65$. This implies a strong penalty to the likelihood of the single category inference (SCI) model. A strict version of the SCI model discussed in the prior literature (e.g. Murphy and Ross (1994)) is thus a poor performer in our task. To give a better chance to the SCI model and account for possible misperception of the categories, we estimated the parameters of each model on a participant-by-participant basis (by maximizing the likelihood)[5].

Table 3.2 reports the mean estimated parameter values (the mean was estimated across participants for whom the focal model is the best). The average parameter estimates are close to the true values for both the rational and the single category model. This suggests that participants used the categories we intended them to

[5]We used box-constraint optimization with 6 free parameters for AM and SCI. The lower bound for all parameters was 0. LM was optimized over 3 parameters without constraints.

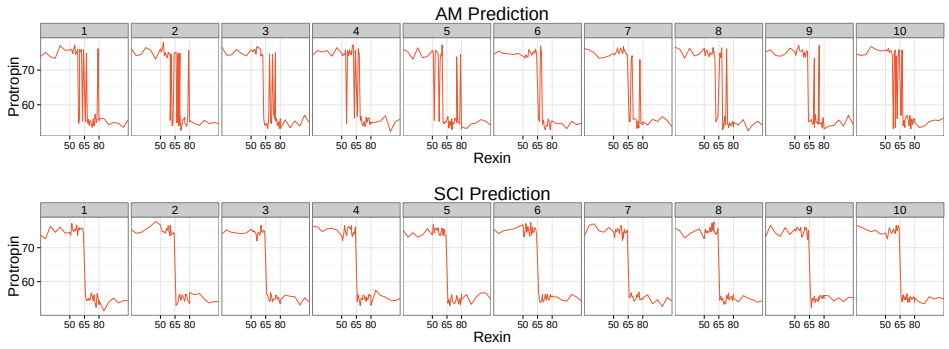


Figure 2.4: Predictions of AM and SCI of inferences of 10 artificial participants. The top row shows predictions of AM, the bottom row shows the predictions of SCI.

use. Models were compared in terms of the BIC criterion. For 60% of the participants, AM provides the best fit while SCI provides the best fit for the rest of the participants.

Analyses of the 'switching' behavior of the participants A crucial prediction of Anderson's rational Model (AM) is that in the area where the 'target' category is uncertain (around the boundary at $x = 65$), there are oscillations between the typical level of Protropin (y-axis) for mice (about 75) and rats (about 55). Consider one participant in the experiment. Suppose the participant has to make an inference about a blood sample with a level of Rexin of 70. The probability that this sample is from a Rat is about 0.8. Anderson's model predicts that in 80% of the cases, a participant facing this situation will give a response close to 55 (typical Protropin level for a Rat sample) and that in about 20% of the cases, she will give a response close to 75 (typical Protropin level for a Mouse sample). If we collect many such judgments in the area where the 'target' category is uncertain, we should expect that some inference values will be close to 55 and others close to 75. The top row of Figure 2.4 shows the inferences of 10 simulated participants who follow Anderson's model. All these simulated participants show oscillations between Protropin levels around 55 and 75 (the x values used for the simulation are the same as those used in the experiments, without any instance of $x = 65$).

By contrast, no such oscillation is implied by the single category model (SCI). In this case, there is just one 'switch' at the boundary ($x = 65$). The bottom row of Figure 2.4 shows the inferences of 10 simulated participants who are assumed to follow the single category model. Inferences are close to 75 for Rexin levels lower than 65 and close to 55 for Rexin levels higher than 65.

Instances of the two distinct inference patterns can clearly be seen on the graphs depicting the inferences of the participants in the experiment (Figure 3.6). For ex-

ample, participant 4 switched exactly once at the 'boundary' whereas participant 8 switched many times between the two modal Protropin levels of 55 and 75. Our participant-by-participant model estimations identified this difference since the single category model provides the best fit to the inferences of Participant 4 whereas Anderson's model provides the best fit to the inferences of Participant 8.

In Experiments 1 there are 12 (40%) participants with exactly one switch. This is very close to the number of participants best fit by the single category model (with parameters estimated participant-by-participant – see Table 1).

2.4.4 Discussion

Most participants' inferences are better explained by Anderson's rational model (AM) than by the single category model (SCI) and the linear model (LM). In this experiment, we provided participants with a visual representation of the categories. One might wonder if this design captures the psychological process that underlies inferences with uncertain categorization when such graphical representation is not available at the time of the inference. It could be that participants engaged in some elaborate form of curve fitting on the basis of the graphs we showed to them. We address this potential concern in Experiments 2 & 3.

2.5 Experiments 2 & 3

The experiments follow a design similar to Experiment 1. We recruited 30 participants via Amazon Mechanical Turk for each experiment.

Design of Experiment 2 The only change in comparison to Experiment 1 is that we removed the graphical representation of the categories (i.e. the graph of Figure 2.1) on the screens on which participants made judgments (during learning and test stages). The graph was shown in the instructions and before every judgment, but not on the judgment screen.

Design of Experiment 3 In this experiment, participants never saw any graphical representation of the data. They learnt the categories from experience by first seeing 40 exemplars of both categories (Rexin and Protropin values), then making within-category inferences (of Protropin level based on Rexin level) with feedback and categorizations of blood samples as Rat or Mouse, based on Rexin level (see Supplementary Material).

2.5.1 Results of Experiments 2 & 3

Parameter-Free Model Comparison Removing the graphical depiction of the data didn't drastically change the pattern of results. Just as in Experiment 1, Anderson's model is the best for the majority of the participants in both experiments (Table 3.1).

Comparison of Models with Parameters Estimated Participant-by-Participant and Switching Behavior In comparisons based on the BIC, Anderson's model (AM) is by far the best fitting model (Table 3.2). As in Experiment 1, the single-category model (SCI) performs better in this comparison than in the comparison of parameter-free models, but worse than Anderson's model.

In Experiments 2 & 3 there are 4 (13%) and 8 (28%) participants with exactly one switch (see also participant-by-participant inferences in Figures S3&S4 in the Supplementary Material). These numbers closely reflect the performance of the single-category model (with parameters estimated participant-by-participant).

2.6 Discussion of Experiments 1-3

Taken together, Exp. 1-3 show that Anderson's model (AM) provides a better fit to the data than the single category model (SCI). The linear model provides a very poor fit to the data. This pattern of results is consistent across experiments, in which we varied the information available about the categories. Whether participants saw a graphical representation of the categories in feature space at time of inference (Exp. 1), this representation was seen in the learning stage but removed in the inference stage (Exp. 2), or never seen (Exp. 3), the patterns of feature inferences were similar.

We would like to claim that good performance of Anderson's model is evidence for the integration of information across categories. In other words, we would like to claim that people use a cognitive algorithm of the following kind:

1. observe $X = x$;
2. compute the posterior distribution $f(y \mid x)$ according to eq. 2.7;
3. provide an estimate of Y by generating a random draw from the posterior.

Because the posterior depends on the marginal distributions of the unobserved features for both the target and the non-target category, we call this cognitive algorithm 'AM non-target'.

The evidence gathered so far does not unequivocally show that participants used this kind of cognitive algorithm. The reason is that the results of Exp. 1-3 are also compatible with a noisy version of the Single Category Inference (SCI) model. Suppose a participant uses SCI, but is uncertain about the location of the boundary at which the 'Rat' category becomes more likely than the 'Mouse' category. Let β denote the uncertain position of the boundary on the x-axis. Suppose inferences about y are produced by the following cognitive algorithm:

1. observe $X = x$ and estimate the position of the boundary β;
2. evaluate if $x < \beta$ or if $x > \beta$.
3. if $x < \beta$, select the 'Mouse' category; else select the 'Rat' category. Denote the selected category by c^*.

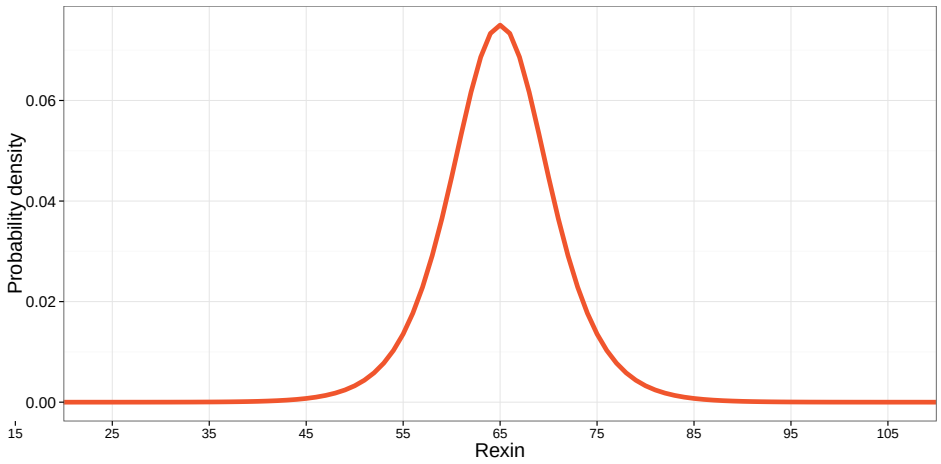


Figure 2.5: Density of the position of the uncertain boundary between the categories 'Rat' and 'Mouse' implied by the second cognitive algorithm (SCI with uncertain boundary).

4. provide an estimate of Y given $X = x$ by producing an intuitive estimate of the mode of the posterior distribution conditional on the selected category $f(y \mid c^*)$.

Assume, moreover, that the uncertainty is such that the participant's belief about the location of this boundary is represented by a probability density $g(\beta) = \frac{\partial P(Rat|\beta)}{\partial \beta}$. In this case, the inferences produced by this algorithm are compatible with Anderson's rational model: $P(Rat \mid \beta)$ follows eq. 3.9, assuming $c_1 = Rat$ and $x = \beta$ – see Supplementary Material for an explicit formulation of $g(\beta)$. Figure 2.5 depicts the density of the uncertain boundary, $g(\beta)$. It is a unimodal symmetric distribution centered at the mid-point between the two categories ($x = 65$). We will refer to this algorithm as 'SCI with uncertain boundary'.

If participants whose inferences are best fit by Anderson's model rely on the 'SCI with uncertain boundary' cognitive algorithm, then eliminating the uncertainty about the boundary should reduce the fit of Anderson's model as compared to SCI. This should not happen if people integrate information from both categories in inferring the unobserved feature value – if they use the 'Anderson non-target' algorithm. We designed an experiment that relied on these predictions.

2.7 Experiment 4

Experiment 4 used a design identical to Exp. 1, with one change: we provided participants with information that ruled out subjective uncertainty about the boundary at which one category becomes more likely than the other. Consider the following

two hypotheses:

- H1: People whose inferences are best fit by Anderson's rational model use the 'SCI with uncertain boundary' cognitive algorithm.
- H2: People whose inferences are best fit by Anderson's rational model use the 'AM non-target' cognitive algorithm.

If hypothesis H1 is true, then removing the subjective uncertainty about the boundary should lead to inferences consistent with the SCI model. Therefore, under this hypothesis, the SCI model should provide a better fit to participants' inferences than in Exp. 1. If hypothesis H2 is true, removing subjective uncertainty about the boundary should not lead to a relative increase in the performance of the SCI model.

2.7.1 Design

31 participants recruited via Amazon Mechanical Turk completed the experiment. The design was identical to Experiment 1 except for the addition of the following note below the graph depicting the blood sample data: "Note: A blood sample with a Rexin level equal to 65 is equally likely to come from a Rat or a Mouse." (Figure S2 in Supplementary Material). The note was shown whenever the graph was shown. We checked that the participants understood the meaning of the note about the boundary in three True-False questions (see Supplementary Material).

2.7.2 Results

Comparison Parameter-Free Models Anderson's model is the best fitting model for 74% of the participants (see also Figure S5 in Supplementary Material). This performance is very similar to that of Exp. 1-3. The performance of Anderson's model is even higher among the 24 participants who passed the comprehension check (AM provides the best fit for 88% of these participants).

Comparison of Models with Parameters Estimated Participant-by-Participant and Switching Behavior Maximum likelihood estimation of the parameters and the analysis of the switching behavior yields results similar to those of Exp. 1-3 (Table 3.2). There are 8 out of 31 participants (26%) with exactly one switch.

2.7.3 Discussion of Experiment 4

Informing the participants of which category was the more likely did not lead to a performance improvement of the single category model (SCI) model relative to Anderson's model (AM). This is inconsistent with Hypothesis H1 but consistent with Hypothesis H2. This suggests that participants who made inferences consistent with Anderson's rational model (in Exp. 1-3) did not use the 'SCI with

uncertain boundary' cognitive algorithm. Rather, it likely reflects the operation of a cognitive algorithm in which participants' inferences were affected not only by the more likely category but also by the non-target category.

2.8 Discussion of Experiments 1-4

The first four experiments analyzed the performance of Anderson's rational model in a task environment characterized by conditional independence. We found that in these four experiments, Anderson's model performed well – unequivocally better than the single category inference (SCI) model. In other words, participants' inferences were influenced by the non-target category. This seems to contradict the findings of prior studies that found evidence in favor of the SCI model.

A skeptic might wonder whether the difference between our results and the results of these studies could be due to the fact that our design differs in many ways from the standard paradigm used to study feature inference under uncertain categorization.[6] After all, our design differs from this paradigm not only in terms of satisfying the conditional independence assumption, but also on other dimensions: continuously valued features versus discrete features, flow of the experiment, number of categories, etc... We believe that relying on a discrete paradigm to study the performance of Anderson's model is suboptimal because in this context the predictions of Anderson's model and the predictions of the feature conjunction model (a model that ignores categories altogether) are the same, as explained in the first section of the paper. Nevertheless, we wanted to assess the performance of Anderson's model in a setup that matched the standard paradigm as closely as possible. This is the purpose of the next experiment.

2.9 Experiment 5

The experiment was designed to fit within the standard discrete feature paradigm while satisfying the conditional independence assumption.

Design

The design of this study closely follows the designs of Experiment 1 and 2 in Murphy and Ross (2010b) and Experiment 1 in Murphy and Ross (2010a). Just as in the original experiments participants were shown drawings by four children (see Figure 2.6 for one of the panels used in the experiment). Each category consisted of nine drawings and each drawing was a colored shape. The stimuli were designed such that in each category they were conditionally independent. For example, in the 'Kyle' category there were 4 orange circles, 2 green circles, 2 orange squares and 1 green square. We have $P(orange\&circle) = 4/9$. We also have $P(orange)P(circle) = 6/9 * 6/9 = 4/9$.

[6] We thank an anonymous reviewer for expressing this skepticism.

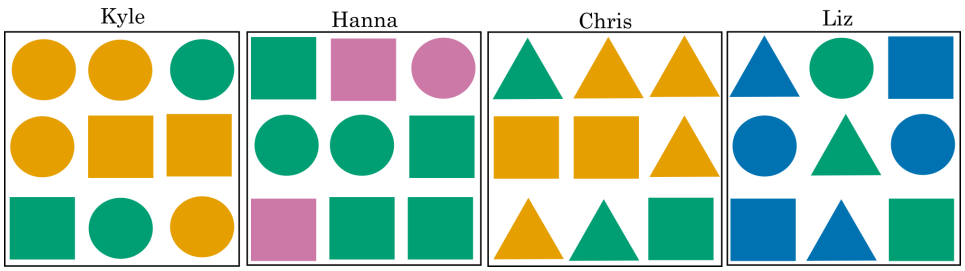


Figure 2.6: One of the panels used in the Experiment 5.

The flow of the experiment was as follows. First, participants were shown general instructions about the children drawings. They were told each collection of 9 children drawings was a sample form a larger set of drawings by this child. Each question had several parts, as shown in the following example:

- I have a new drawing that is a circle.
- What color do you think this circle drawing has?
- What is the probability (0-100%) that it is has this color?

The instructions also explained that 0 on the scale means that it is impossible, 50% means that this happens about half of the time, 100% means that this is completely certain. Participants answered four such questions, two about one panel, and two about another similar panel. For each panel, one question used shape as a given feature and the other one used color.

In most of the original studies (e.g. Experiment 1 in Murphy & Ross, 2010b), participants were asked about the target category after being informed of the feature (e.g., that the shape was a circle) and before making their prediction (e.g. about the color of the circle). In the above example, participants would be asked "Which child do you think drew it?" after reading "I have a new drawing that is a circle". They were also asked for their confidence in that judgment ("What is the probability (0-100%) that the child you just named drew this?"). Because prior research suggests that this question makes people more likely to rely on the SCI strategy (e.g., Murphy & Ross, 2010b), we created three between-subject conditions. In the 'NO' condition, participants were not asked asked about the target category. In the 'PR' condition, participants used four sliders to indicate the probabilities that each child had drawn the shape (the sum of the 4 probabilities was constrained to be 100). In the 'MC' condition, participants answered a multiple-choice question (with the 4 children as possible answers) about the target category and their confidence in that judgment, as described at the beginning of this paragraph.

In order to closely match the design of prior experiments, there were two within-subject conditions. In the *agree* condition, the predictions of the SCI and

the AM models were the same. In the *disagree* condition, the predictions of the SCI and AM models differed. For example, with the panel of Figure 2.6 the question in the *agree* condition was about the shape of a green drawing. In this case, both models predict that the response is 'square'. The question in the *disagree* condition was about the color of a 'circle'. Here SCI predicts 'orange' (with 'Kyle' as the target category) whereas AM predicts 'green'. See Table 2.3 for the details on the models' predictions.

92 participants recruited via Amazon Mechanical Turk completed the experiment for a flat participation fee. There were 31 participants in the NO condition, 30 participants in the PR condition and 31 participants in the MC condition.

Table 2.3: Predictions of the models for the color of a circle for panel on Figure 2.6. AM: Anderson's Model, SCI: Single Category Independent Features Model, LM: Linear Model or Feature Conjunction Model.

	Model		
Response	AM	SCI	LM
P(orange | circle)	P(Kyle | circle)*P(orange | Kyle)+P(Hanna | circle)*P(orange | Hanna) +P(Ch | circle)P(orange | Ch)+P(Liz | circle)*P(orange | Liz)= =6/12*6/9+3/12*0+0*6/9+3/12*0=4/12	P(orange | Kyle)= **6/9**	=4/12
P(green | circle)	P(Kyle | circle)*P(green | Kyle)+P(Hanna | circle)*P(green | Hanna) +P(Ch | circle)P(green | Ch)+P(Liz | circle)*P(green | Liz)= =6/12*3/9+3/12*6/9+0*3/9+3/12*3/9= **5/12**	P(green | Kyle)= 3/9	**=5/12**
P(purple | circle)	P(Kyle | circle)*P(purple | Kyle)+P(Hanna | circle)*P(purple | Hanna) +P(Ch | circle)P(purple | Ch)+P(Liz | circle)*P(purple | Liz)= =6/12*0+3/12*3/9+0*0+3/12*0=1/12	P(purple | Kyle)= 0	=1/12
P(blue | circle)	P(Kyle | circle)*P(blue | Kyle)+P(Hanna | circle)*P(blue | Hanna) +P(Ch | circle)P(blue | Ch)+P(Liz | C)*P(blue | Liz)= =6/12*0+3/12*0+0*0+3/12*6/9=2/12	P(blue | Kyle) = 0	=2/12

Table 2.4: Results of Experiment 5. Percentage of participants who's feature inferences corresponded to the models' predictions. For proportion comparison value of Person's χ^2 test statistic and confidence interval of the proportion difference are reported.

	Condition		
Model	NO	PR	MC
AM: Anderson	40 (65%)	32 (53%)	22 (35%)
SCI: Single Cat. Indep. Features	12 (19%)	18 (30%)	33 (53%)
χ^2	24.1	5.8	3.3
95CI	[0.28; 0.62]	[0.05; 0.42]	[-0.36; 0.01]
# observations	62	60	62
# participants	31	30	31

Results

Unsurprisingly, the majority of participants' responses for the *agree* condition questions were consistent with the predictions of AM and SCI (which are the same). The proportion of predicted responses were 61%, 67% and 76% in the NO, PR and MC conditions, respectively. (Each participant answered one question in both *agree* and *disagree* conditions for each panel; the resulting numbers of observations are thus 62, 60 and 62 for NO, PR and MC conditions respectively.)

We now turn to the results of the *disagree* condition questions. For each condition we calculated the proportion of participants' responses that correspond to the predictions of the AM and SCI models. In the NO and PR conditions, participants' inferences were much more consistent with Anderson's model than with the single category model. The response predicted by AM was chosen 65% and 53% of time in the NO and PR conditions, respectively. By contrast, the response predicted by SCI was chosen much less frequently at 19% and 30% of the time, respectively. The differences in proportions are significant (see Table 2.4 for details).

The results differ in the MC condition. In this case, SCI provides a much better fit to participants' inferences. The response predicted by SCI was chosen 53% of the time, whereas the response predicted by AM was chosen 35% of the time.

We find that in the "PR" and "NO" conditions the answer that corresponds to AM prediction was chosen significantly more (by 53% and 65% of the participants respectively, see Table 2.4 for details). In the "MC" condition an answer that corresponds to SCI prediction was chosen by 53% of the participants. This is marginally significantly (at the 10% level) higher than the proportion of answers that corresponds to AM prediction.

Discussion

The purpose of this experiment was to evaluate the performance of Anderson's rational model in an environment with discrete features and conditional independence. We found that Anderson's model performs well provided that participants were not asked to categorize the shape before formulating their feature inference. This suggests that participants were able to take into account more than just the target category, as in Exp. 1 to 4. Whether they considered the various candidate categories and weighted them optimally (or close to optimally) or they ignored categories altogether cannot be decided on the basis of these data, because the prediction of Anderson's model (optimal weighting of the candidate categories) and of the feature conjunction strategy (ignoring the categories) are the same. Taken together with the results of Exp. 1 to 4, however, the results of this experiment suggest that participants were sensitive to categories and were influenced by more than just the target category when formulating their feature inferences.

Although our focus here is not on how categorization questions affect inferences, it is worth discussing our results about the three between-subject conditions (NO, PR and MC conditions). We found that when participants were asked to select the most likely category of the object ('MC' condition'), the single category inference model was the best fitting model. In the other conditions, Anderson's model was the best fitting model. This suggests that the question format strongly affects the extent to which feature inferences are influenced by one or several categories. These are consistent with the results of the experiments reported by Murphy and Ross (2010b). In their experiments, they found that when participants were asked to evaluate the probabilities of the four candidate categories, their feature inferences were influenced not only by the target category but also by the other candidate categories (the design of their Experiment 1 is closest to our MC condition and the design of their Experiment 2 is closest to our PR condition).

Related findings were reported by Hayes and Newell (2009, Exp. 3) and Murphy and Ross (1994, Exps. 5&6). In these experiments, the authors compared feature inference between a setting in which participants had to categorize items before the feature prediction (this condition is closest to our MC condition) and a setting where they did not have to make such categorization decision (this condition is closest to our NO condition). Both of the earlier studies found that inferences were more likely to be influenced by several categories when they did *not* follow a categorization decision. This finding is similar to the comparison of our MC and NO conditions: in the NO condition, inferences were much more likely to be influenced by several categories than in the MC condition.

Taken together the comparisons of inferences between the MC condition on one side and the PR and NO conditions on the other side are consistent with a conjecture according to which much of the past evidence for the effect of a single category on feature inference could be due to the ordering of the questions in most experiments (categorization first and feature inference second). A comparison of inferences in the NO and PR conditions possibly supports this conjecture. In the

NO condition, feature inferences were similar to those in the PR condition. This suggests that participants who were *not made to focus on just one category* by a categorization question readily consider more than just the target category (at least in the setting of our experiment, where the target category was not immediately clear). This in turn suggests that people's default strategy when the category is uncertain is to consider more than one category. This default tendency could be overridden by having people focus on just one category, but more research is clearly needed to evaluate this conjecture.

2.10 General Discussion

2.10.1 Integration of information over categories

This paper contributes to the literature that addresses the extent to which several categories affect feature inference when categorization is uncertain. In our studies participants' inferences were affected by several categories, which suggests that they integrated information across these categories. But what does it mean that participants *integrate* information across categories?

Information integration can happen at two levels in the model: when computing the posterior distribution and at the level of the decision rule. In our rendition of Anderson's model, information integration occurs in the computation of the posterior distribution. To see this, note that the posterior is bimodal when the more likely category is uncertain. Importantly, in this model, there is no information integration at the level of the decision rule. We assumed that the inferred feature value was a random draw from the posterior distribution. It is possible to think of alternative models where information integration also operates at the level of the decision rule. Maybe the most straightforward choice for such a decision rule is the expected value of the posterior $E[Y \mid x]$, computed with respect to the posterior produced by Anderson's model. Under conditional independence this is:

$$E[Y \mid x] = P(c_1 \mid x)\mu_{y1} + P(c_2 \mid x)\mu_{y2} = P(c_1 \mid x)(\mu_{y1} - \mu_{y2}) + \mu_{y2}. \qquad (2.10)$$

In this case, the inferred feature value is the weighted average of the means of the categories. This 'AM Averaging' model is sensitive to the overall feature correlation, like the linear model we estimated on our data (LM), but it is more sophisticated: the mean of the posterior looks like a logistic curve (see Fig. 1 in Konovalova and Le Mens (2016) for an example). Such response curve predicts that in the area of uncertainty (x around 65) subjects would give y values that lie in between the category means (μ_{y1} and μ_{y2}). More generally, response models that rely on some form of averaging of the category-specific inferences to produce an inference in the area where the category is uncertain will produce a unimodal conditional distribution of response. In other words, if $g(y \mid x)$ denotes the conditional density of responses obtained for a specific x values, a model that relies on

averaging will be such that $g(\cdot \mid x)$ is a unimodal distribution[7]. This prediction is inconsistent with the evidence we obtained in our four experiments. Examination of the participant-by-participant data clearly shows that responses are bimodal in the area where the category is uncertain (x close to 65 — see Fig. 3.6 and Figs. S3-S5 in the Supplementary Material).

To illustrate this further, we assessed the performance of the 'AM Averaging' model on our data. We specified the 'AM Averaging' model with a normally distributed error term (to account for the likely dispersion of responses around the deterministic prediction produced by human error). This results in the following posterior:

$$f(y \mid x) = f_{E[Y|x],\sigma_{AMA}}(y), \tag{2.11}$$

where $f_{E[Y|x],\sigma_{AMA}}(y)$ denotes a normal *pdf* with a mean $E[Y \mid x]$ (given by eq. 2.10) and standard deviation σ_{AMA}.[8] In comparisons of the AM, SCI and 'AM Averaging' models, we found that the 'AM Averaging' model provides the best fit to almost exactly the same number of participants as the linear model in the analyses reported above (see Table S1 in the Supplementary Material). In all cases, the number of participants best fit by the 'AM Averaging' model is substantially lower than the number of participants best fit by Anderson's model (AM).

This analysis suggests that most participants integrated information from the candidate categories in computing the posterior, but not in their decision rule. More work is clearly needed to understand which type of decision rule best explains feature inferences in the kind of continuous environment we have used in Exp. 1 to 4. Other decision rules are possible. For example, one could think of a decision rule that selects the mode of the posterior distribution, or a decision rule that consists of a random draw from a distribution that is centered around the mode. A possible way to study decision rules is to specify an explicit reward function. The form of the reward function will likely affect the decision rule used to formulate the inferences. A related question pertains to how the task environment affects the decision rule when no reward function is explicitly specified.

2.10.2 Relation to Nosofsky's exemplar model

In a recent paper, Nosofsky (Nosofsky, 2015) proposed an exemplar model that provides a good fit to existing data based on the discrete paradigm just discussed. Like ours, this model makes feature inferences that are influenced by all the categories. And for most parameter values, the target category receives a higher weight.

[7] Unless one assumes a very bizarre error term distribution.

[8] In the parameter-free version of the model, we took $\sigma_{AMA} = 4.7$. This value was obtained by maximum likelihood estimation of the model (eq. 2.11) with just σ_{AMA} as a free parameter and the true values for the other parameters ($\mu_{xR} = 80, \mu_{yR} = 55, \mu_{xM} = 50, \mu_{yM} = 75, \sigma_{xR} = \sigma_{xM} = 10, \sigma_{yR} = \sigma_{yM} = 1$), based on all the exemplars depicted on Figure 2.1 (irrespective of their categories). In the version of the model with parameters estimated on a participant-by-participant basis, σ_{AMA} and all the other parameters were estimated by maximum likelihood, just as in the analyses of the experiments reported in earlier sections.

The model relies on an assessment of the similarity between the observed stimulus and the data stored in memory and then does some similarity weighted prediction. The way the similarity is computed gives more weight to the most likely category (the 'target' category). It also gives some weight to the other categories. Therefore, just like our model, the exemplar model makes inferences that are influenced by several categories. This model differs from ours because the exemplar model is an algorithmic model whereas our model is specified at the computational level. Ours does not specify the details of the mental computations whereas the exemplar model does.

To cast light on the relation between Nosofsky's model and the other models discussed in this paper, we computed the predictions of Nosofsky's model in the context of the Experiment 5 reported earlier. The model has a parameter S, that characterizes the weight of exemplars that do not have the observed feature. The second parameter, L, regulates the sensitivity to the target category. Intuitively, when $L = 0$ the individual pays no attention to exemplars outside the target category. Let $\hat{P}(y \mid x)$ denote the probability that the second feature is equal to y if the first feature is observed to be equal to x according to the exemplar model (we use the 'hat' to emphasize that this is the model prediction). When $L = 0$, the model predicts:

$$\begin{aligned}
\hat{P}(orange \mid circle) &= \frac{6}{9} \\
\hat{P}(green \mid circle) &= \frac{3}{9} \\
\hat{P}(purple \mid circle) &= 0 \\
\hat{P}(blue \mid circle) &= 0
\end{aligned}$$

These are the same predictions as SCI (see Table 2.3). When $L = 1$ and $S = 0$, the exemplar model predicts

$$\begin{aligned}
\hat{P}(orange \mid circle) &= \frac{4}{12} \\
\hat{P}(green \mid circle) &= \frac{5}{12} \\
\hat{P}(purple \mid circle) &= \frac{1}{12} \\
\hat{P}(blue \mid circle) &= \frac{2}{12}
\end{aligned}$$

These are the same predictions as AM (see Table 2.3).

Future research should go beyond this specific case and clarify the exact links between the exemplar model and other computational models of feature inferences both in discrete and continuous environments. Shi et al. (2010) have shown that exemplar models can be seen as algorithms for performing Bayesian inference provided the decision rule can be specified as the expectation of a function. This suggests that it might be possible to show that an exemplar model could approximate

the prediction of the Anderson Averaging model discussed in the prior subsection (note that this model does not fit our experimental data well, however). It might be harder to design an exemplar model that can produce the same predictions as Anderson's model with a decision rule that consists in a random draw from the posterior. This is because this decision rule cannot easily be specified as the expectation of a function.

2.11 Conclusion

Taken together, our results show that in an environment characterized by conditional independence, Anderson's rational model is a good predictor of participants' inferences – a performance much higher than what was suggested by previous empirical research on inference with uncertain categorization. This good performance suggests that most participants were influenced not only by the most likely category (given the observed feature) but also by the other candidate category. There was heterogeneity among participants: a non-trivial minority of participants was best fit by the single-category model. This aspect of our results is consistent with prior research which found heterogeneity in the propensity to rely on just one category - even though this research often found that most participants relied on the most likely category whereas found that most participants relied on multiple categories (Murphy & Ross, 2010a).

Finally, it is important to note that our results do not speak of the realism of the conditional independence assumption. As explained in the introduction, there are many environments where conditional independence does not hold. We are currently adapting Anderson's model to environments with positive or negative within-category correlations. The results reported here suggest that such a model will perform well. Preliminary evidence suggests that it is the case.

2.12 Appendix

2.12.1 Additional Methodological Details

The four experiments had the same structure. After reading general instructions, the participants went through two stages: training and testing. The purpose of the training stage was to teach the participants about the (artificial) categories used for feature inference. The training stage was divided into 4 parts (see Figure 2.7):

- *Learning about levels of Rexin and Protropin for the category 'Rat'.* First, the participants were told that their lab has a collection of Rat blood samples. They were provided with information about the levels of Rexin and Protropin in these blood samples (how this information was communicated differed across experiments – see below for details). To ensure that participants payed adequate attention to the data, they were asked to make a series of inferences of the likely level of Protropin given the level of Rexin found in a 'Rat' blood sample. After making inferences the participants were presented with feedback that consisted of the Rexin level, their answer, and correct level of Protropin.

- *Learning about the levels of Rexin and Protropin for the category 'Mouse'.* The procedure is identical to the 'Rat' category.

- *Learning about the relative positions of the categories in feature space.* First, participants were told that a batch of new blood samples had just arrived at their lab and that these blood samples had already been tested for Rexin. They were also told that the 'label on the blood sample has been erased and thus you do not know if it belongs to a rat or a mouse.' Then they were asked to categorize newly arrived samples as 'Rat' or 'Mouse' based on the level of Rexin in those samples. The next screen presented feedback with the correct information about the sample.

- *Repetition of all three types of tasks.* Each loop consisted of three predictions. First, they made an inference about a Rat sample's level of Protropin, then an inference about a Mouse sample's level of Protropin, and finally, they categorized a sample as 'Rat' or 'Mouse' based on the level of Rexin.

In the testing stage the participants were asked to make inferences about Protropin levels of newly arrived blood samples 'without labels' based on the level of 'Rexin'.

The four experiments differ in terms of the availability of information about the two categories during the training stage (Exp. 1-3) and the testing stage (Exp. 4 vs Exp. 1). We detail the differences in the following.

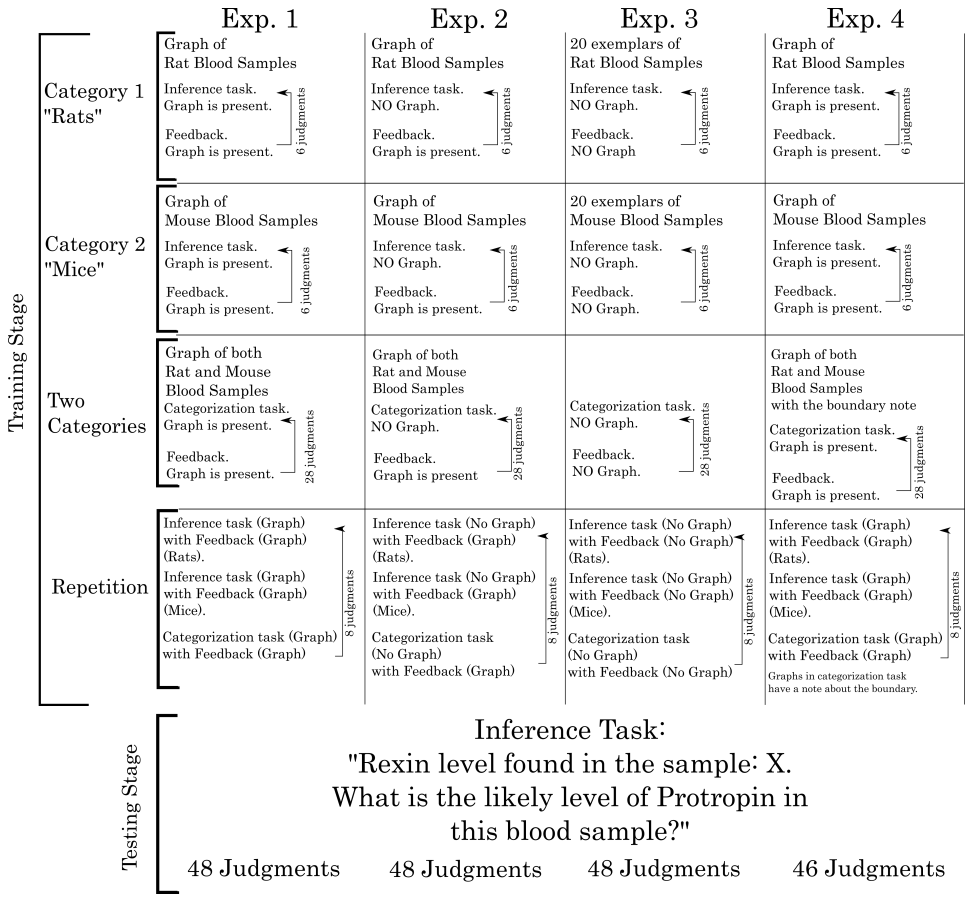


Figure 2.7: Structure of the experiments.

Experiment 1

In Experiment 1, the participants had access to a graphical representation of the data including the screen on which they made their inferences (see column 1 in Figure 2.7). At the beginning of each part of the training stage the participants were shown a scatter plot of the levels of Rexin and Protropin in a large set of blood samples. This scatter plot remained visible during the judgments and on the feedback screen. That is, when learning about the 'Rat' category, participants were shown a scatter plot of the levels of Rexin and Protropin in a large number of 'Rat' blood samples (the part of the graph that pertains to 'Rat' blood samples in Figure 1 in the body of the paper). Similarly, when learning about the 'Mouse' category, participants were shown a scatter plot of the levels of Rexin and Protropin in a large number of 'Mouse' blood samples. In the categorization training stage, participants were shown the graph of Figure 1 in the body of the paper (a scatter plot of the levels of Rexin and Protropin in a large number of blood samples from

the two categories). Similar information was shown at each loop of the fourth, combined, training stage.

In the test stage, the scatter plot of Figure 1 in the body of the paper was visible on the inference screens.

Experiment 2

The only change with comparison to Experiment 1 is that the scatter plots were *not* displayed in any of the judgment screens (feature inferences and categorizations in training stage, and feature inferences in test stage). The scatter plots were shown in all other screens of the training stage (see details in column 2 in Figure 2.7).

Experiment 3

In Experiment 3 participants did not see any graphical representation of the data. Rather, participants learned about the categories purely by experience. Instead of a scatter plot, participants were shown, in a sequential fashion, a representative sample of the data (numeric levels of Rexin and Protropin). Before the first two substages of the training stage, participants were shown 20 exemplars of each category (see details in column 3 in Figure 2.7). Samples were shown in an auto-advancing loop each for 1 second.

Experiment 4

The structure of Experiment 4 was almost identical to Experiment 1 (see details in column 4 in Figure 2.7).[9] In order to decrease uncertainty about the 'boundary' we added a note to the graph depicting both the 'Rat' samples and 'Mouse' samples. The exact text of the note is: 'A blood sample with a Rexin level equal to 65 is equally likely to come from a Rat or a Mouse.' (see Figure 2.8)

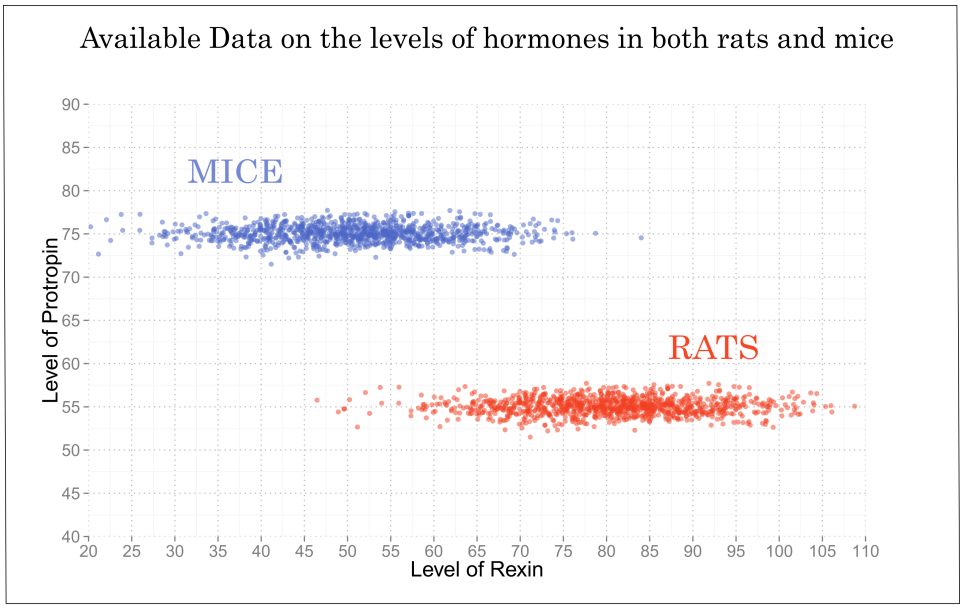
Furthermore, in order to check that participant understood the note we added a block with comprehension questions. It consisted of three questions with binary choices of 'Yes' and 'No":

1. 'Is the following statement true: A blood sample with a Rexin level equal to 65 is equally likely to come from a Rat or a Mouse.'

2. 'Is the following statement true: A blood sample with a Rexin level higher than 65 is more likely to come from a Rat than a Mouse.'

3. 'Is the following statement true: A blood sample with a Rexin level lower than 65 is more likely to come from a Mouse than a Rat.'

[9] A minor difference is that we eliminated the stimuli levels where Rexin was 65 during the testing stage – the reason being that the SCI model does not make any explicit prediction for this x value. As a result participants made 46 judgments instead of 48.



Note: A blood sample with a Rexin level equal to 65 is equally likely to come from a Rat or a Mouse.

Figure 2.8: Graphical depiction of the categories used in Experiment 4.

2.12.2 Participant-Level Inferences

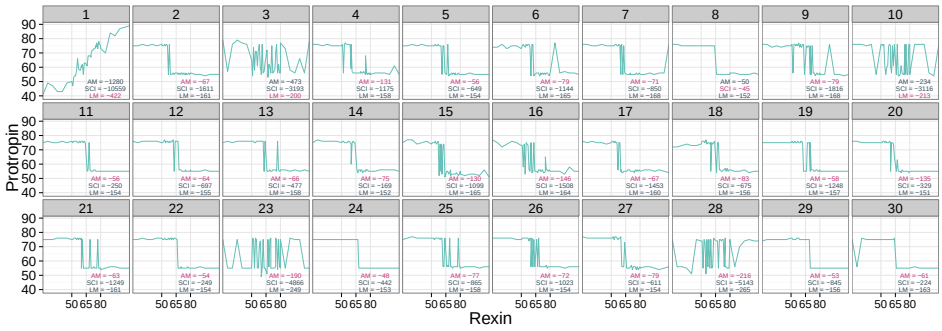


Figure 2.9: Inference of the participants of Experiment 2. The log-likelihoods of each parameter-free model is shown on each participant's graph. AM: Anderson's rational model, SCI: single category independent feature model; LM: linear model. The red font indicates the best fitting model.

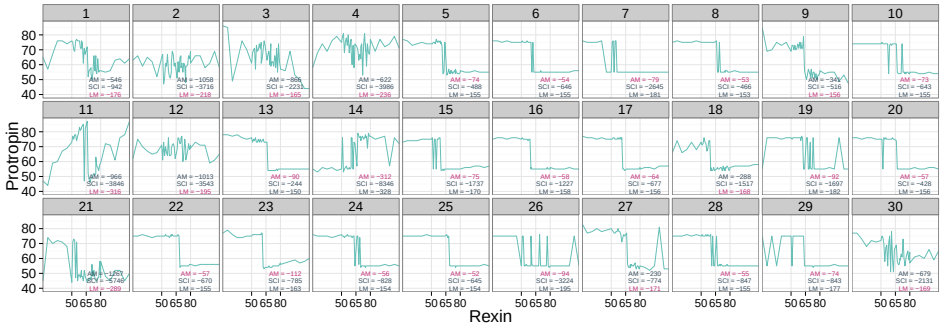


Figure 2.10: Inference of the participants of Experiment 3. The log-likelihoods of each parameter-free model is shown on each participant's graph. AM: Anderson's rational model, SCI: single category independent feature model; LM: linear model. The red font indicates the best fitting model.

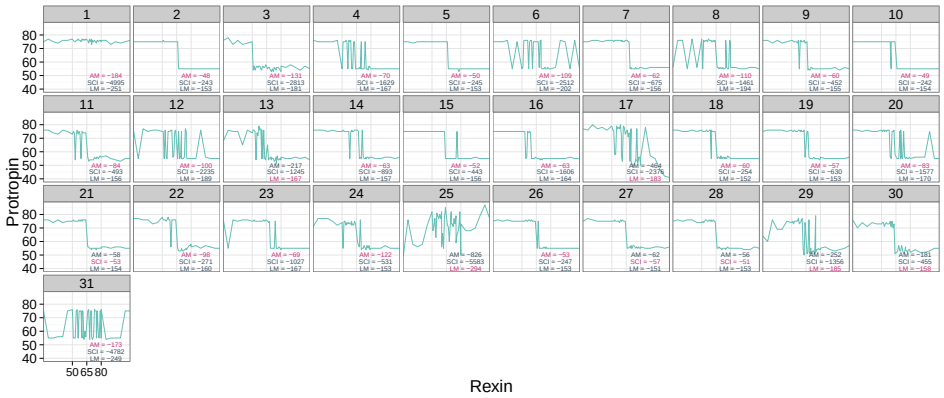


Figure 2.11: Inference of the participants of Experiment 4. The log-likelihoods of each parameter-free model is shown on each participant's graph. AM: Anderson's rational model, SCI: single category independent feature model; LM: linear model. The red font indicates the best fitting model. Participants 25 to 31 are the ones that did not pass the comprehension check about the 'boundary' information.

2.12.3 Note on uncertainty about the position of the 'boundary'

In the Discussion of Experiments 1-3 we noted that it is possible that a participant might be uncertain about the position of the 'boundary' at which the two categories are equally likely. Let β denote the subjective position of the boundary on the x axis and let g denote its probability density function (*pdf*).

Suppose a participant uses the single category model (SCI) with β as the boundary (the second cognitive algorithm in the discussion of the body of the paper). The participant assumes that the blood sample comes from a Rat whenever she believes $\beta < x$. We thus have $P(\beta < x \mid x) = P(Rat \mid x)$.

Under the assumption that this cognitive algorithm produces the same pattern of inferences as Anderson's rational model, the right hand side is given by eq. 8 in the body of the paper (by substituting *Rat* for c_1). This term can be seen as the cumulative distribution function (*cdf*) of β. Therefore, we can easily deduce the implied density (*pdf*) of β:

$$g(x) = \frac{\partial P(Rat|x)}{\partial x} = -(2ax-b)\frac{e^{ax^2-bx+c}}{(1+e^{ax^2-bx+c})^2}$$

with

$$\begin{aligned}
a &= \frac{\sigma_{xM}^2 - \sigma_{xR}^2}{2\sigma_{xM}^2\sigma_{xR}^2}, \\
b &= \frac{\sigma_{xM}^2\mu_{xR} - \sigma_{xR}^2\mu_{xM}}{\sigma_{xM}^2\sigma_{xR}^2}, \\
c &= \frac{\sigma_{xM}^2\mu_{xR}^2 - \sigma_{xR}^2\mu_{xM}^2}{2\sigma_{xM}^2\sigma_{xR}^2} + \log\frac{\sigma_{xM}}{\sigma_{xR}} + \log\frac{P(Mouse)}{P(Rat)}.
\end{aligned}$$

2.12.4 Experiment 5 - Additional Analyses

In Experiment 5 we replicated the design developed by Murphy and Ross (2010b). There were two panels portraying four categories with children drawings (Figure 6 in the body of the paper and Figure 2.12 in the Supplementary Material). Following the original design, there were two tasks: an "agree" task and a "disagree" task. In the "agree" task the predictions of the SCI and AM were the same. In the "disagree" task the predictions differed. For example an "agree" task for Figure 2.12 would be about shape of a black figure. Both models predict "circle". A "disagree" task would be about color of triangle. According to SCI the answer should be "orange", whereas AM predicts "black".

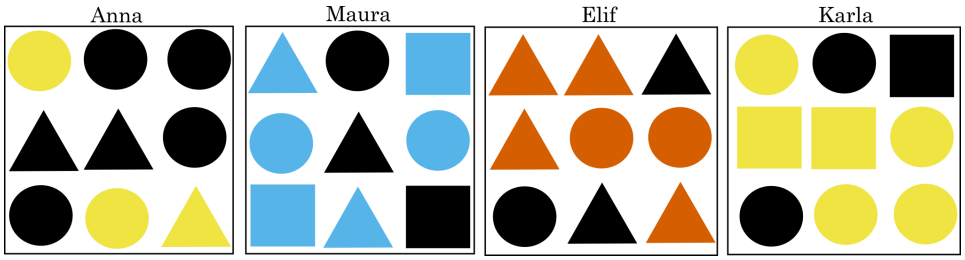


Figure 2.12: Second panel of objects used in the Experiment 5. An "agree" task would be about shape of a black figure. A "disagree" task would be about the color of a triangle.

2.12.5 Performance of the 'AM Averaging' model

We report two sets of analyses, similarly to what we did in reporting the experimental results for Exp.1-4 in the body of the paper. First, we compare the parameter-free models. Second, we compare the models with parameters estimated from the participants' inferences. We computed the log-likelihoods on a participant-by-participant basis. Table 2.5 is similar to Table 3.1 in the body of the paper but here we substitute the linear model (LM) by the AM Averaging (AMA) model. It reports the comparison between three candidate models: AM, SCI and AMA. The performance of the AM Averaging model is very close to the performance of the linear model reported in the body of the paper. We also examined what happens when all 4 candidate models are compared against each other (Table 2.6). The AM Averaging performs better than the linear model (LM), but AM is still the best performer by far: it provides the best feet for the majority of participants in all experiments.

Table 2.5: Percentage of participants whose feature inferences were best fit by each of the candidate models.

	Exp. 1		**Exp. 2**		**Exp. 3**		**Exp. 4**	
Model	True	Est.	True	Est.	True	Est.	True	Est.
AM: Anderson	22(74%)	18(60%)	26(87%)	26(87%)	19(63%)	20(67%)	23(74%)	22(71%)
SCI: Single Cat. Indep. Features	4(13%)	11(37%)	1(3%)	3(10%)	0	9(30%)	3(10%)	9(29%)
AMA: AM Averaging	4(13%)	1(3%)	3(10%)	1(3%)	11(37%)	1(3%)	5(16%)	0
# participants	30		30		30		31	

Table 2.6: Percentage of participants whose feature inferences were best fit by each of the candidate models.

	Exp. 1		**Exp. 2**		**Exp. 3**		**Exp. 4**	
Model	True	Est.	True	Est.	True	Est.	True	Est.
AM: Anderson	22(74%)	18(60%)	25(83%)	26(87%)	19(63%)	17(57%)	23(74%)	21(68%)
SCI: Single Cat. Indep. Features	4(13%)	11(37%)	1(3%)	3(10%)	0	9(30%)	3(10%)	8(26%)
LM: Linear	1(3%)	0	3 (10%)	0	6(20%)	4(13%)	2(6%)	2(6%)
AMA: AM Averaging	3(10%)	1(3%)	1(3%)	1(3%)	5(17%)	0	3(10%)	0
# participants	30		30		30		31	

2.12.6 Analysis of Nosofsky's exemplar model

The model has two parameters, S and L that regulate the weight of exemplars in computations of the posterior probabilities of the second feature. S regulates the sensitivity to the differences in feature values. It is equal to 1 if objects match on the target dimension and $S \in [0,1]$ if they mismatch. L regulates the sensitivity to the target category. It is equal to 1 if objects belong to the same category and $L \in [0,1]$ if they do not.

Let $\hat{P}(y \mid x)$ denote the posterior, according to Nosofsky's model. For example, consider predictions of the exemplar model for the color of a circle (see display on the Figure 2.6). We have:

$$\hat{P}(orange \mid circle) = \frac{4+2S+6SL}{(6+3S)+(6SL+3L)+(9SL)+(6SL+3L)} \tag{2.12}$$

$$\hat{P}(green \mid circle) = \frac{2+S+3L+9SL}{(6+3S)+(6SL+3L)+(9SL)+(6SL+3L)} \tag{2.13}$$

$$\hat{P}(purple \mid circle) = \frac{L+2SL}{(6+3S)+(6SL+3L)+(9SL)+(6SL+3L)} \tag{2.14}$$

$$\hat{P}(blue \mid circle) = \frac{2L+4SL}{(6+3S)+(6SL+3L)+(9SL)+(6SL+3L)} \tag{2.15}$$

Suppose $L = 0$. We have:

$$\hat{P}(orange \mid circle) = \frac{4+2S}{(6+3S)} = \frac{2}{3} = \frac{6}{9} \tag{2.16}$$

$$\hat{P}(green \mid circle) = \frac{2+S}{(6+3S)} = \frac{1}{3} = \frac{3}{9} \tag{2.17}$$

$$\hat{P}(purple \mid circle) = \frac{0}{(6+3S)} = 0 \tag{2.18}$$

$$\hat{P}(blue \mid circle) = \frac{0}{(6+3S)} = 0 \tag{2.19}$$

$$\tag{2.20}$$

These are the same predictions as SCI (see 2.3).

Suppose $L = 1$. We have:

$$\hat{P}(orange \mid circle) = \frac{4+2S+6S}{(6+3S)+(6S+3)+(9S)+(6S+3)} = \frac{4+8S}{12+24S} = \frac{4}{12} \tag{2.21}$$

$$\hat{P}(green \mid circle) = \frac{2+S+3+9S}{(6+3S)+(6S+3)+(9S)+(6S+3)} = \frac{5+10S}{12+24S} \tag{2.22}$$

$$\hat{P}(purple \mid circle) = \frac{1+2S}{(6+3S)+(6S+3)+(9S)+(6S+3)} = \frac{1+2S}{12+24S} = \frac{1}{12} \tag{2.23}$$

$$\hat{P}(blue \mid circle) = \frac{2+4S}{(6+3S)+(6S+3)+(9S)+(6S+3)} = \frac{2+4S}{12+24S} \tag{2.24}$$

If, in addition, $S = 0$, we have

$$\hat{P}(orange \mid circle) = \frac{4}{12} \tag{2.25}$$

$$\hat{P}(green \mid circle) = \frac{5}{12} \tag{2.26}$$

$$\hat{P}(purple \mid circle) = \frac{1+2S}{12+24S} = \frac{1}{12} \tag{2.27}$$

$$\hat{P}(blue \mid circle) = \frac{2+4S}{12+24S} = \frac{2}{12} \tag{2.28}$$

These predictions are the same predictions as AM (see Table 2.3).

Chapter 3

A RATIONAL ANALYSIS OF INFERENCES WITH UNCERTAIN CATEGORIZATION

Joint with Gaël Le Mens

3.1 Introduction

In Chapter 2, we argued that the failure of Anderson's rational model to account for the inferences with uncertain categorization stems from the inconsistency between the task environment and the core assumptions of the model. In particular, we noted that the studies reported or analyzed in existing research relied on task environments without conditional independence. The poor performance of the model in these environments suggests that people do not make such assumption. But it does not imply that participants' inferences are inconsistent with the principles of probability calculus, despite the fact that Anderson's model is sometimes referred to as the 'rational model'. The conditional independence assumption is the model's core assumption about the representation of the mental categories, therefore, in settings with within-category feature correlations, Anderson's model cannot be seen as the rational model. However, the findings reported in Chapter 2 illustrate that when the conditional assumption is satisfied, Anderson's model performs well both in discrete and continuous environments. This result indicates that an extension of this model that relaxes the conditional independence assumption might be a good candidate for explaining human inferences when the conditional independence is violated.

In this chapter, we propose an extension of Anderson's model that allows for the mental categories with non-zero feature correlation. The model substitutes the

marginal probability conditional on the category, $f(y \mid c)$ in the eq. 2.1 in Section 2.1 for the marginal probability conditional on the category and *the observed feature*, $f(y \mid cx)$. The inferences made by our model strictly follow the rules of probability calculus. Thus, our model makes rational predictions (given the constraints imposed by the mental representation of the categories). By contrast, Anderson's model is consistent with the law of total probability only if it is correct to assume conditional independence – that is, only if conditional independence holds in the environment. Whenever this assumption does not hold, the computations that lead to the posterior $f(y \mid x)$ are inconsistent with the rules of probability. Anderson's model is thus cannot be seen as a rational model.

This simple modification has important implications. First, it considerably expands the relevance of the rational model. Although some categories like 'freely interbreeding species' (Anderson, 1991, p. 411) may be characterized by independence, there are many environments in which this assumption is not satisfied (Murphy & Ross, 2010a). For example, a simple co-occurrence of features in categories can result in the within-category correlation. For instance, male deer tend to be bigger and have different coloration than females (Murphy & Ross, 2010a, p. 14) or South East Asian restaurants tend to have tofu and rice on the menu. There is also evidence that people store such correlations in their semantic memory (McRae, Cree, Westmacott, & De Sa, 1999) and use them in categorization and inference tasks (Malt & Smith, 1984; Anderson & Fincham, 1996; Wattenmaker, 1991; Crawford, Huttenlocher, & Hedges, 2006). Besides, in virtually all the settings where people believe that there is a causal relationship between two variables (e.g. educational achievement and income, quality and price of consumer goods), the corresponding mental representation invokes a within-category correlation (Rehder & Hastie, 2004).

Second, relaxing the conditional independence assumption allows for a reinterpretation of existing findings. We are able to make sense of several patterns found in experiments on feature prediction with uncertain categorization: the good performance of the feature conjunction strategy that relies on the overall feature correlation and ignores the categories and the good performance of a version of the exemplar model proposed by Nosofsky (2015).

The first point refers to the fact that when features are discrete, our rational model makes the same predictions as the (multiple-category) 'feature conjunction' (see Section 2.2 for more details). This strategy ignores the categories and only relies on the feature conjunctions. It was found to provide the best fit to most existing data (Murphy & Ross, 2010a; Hayes et al., 2007; O. Griffiths et al., 2012, 2011; Papadopoulos et al., 2011). Our re-analysis of the existing findings suggests a different interpretation of this success. As the rational model makes the exact same prediction, one cannot rule out that participants' inferences were, in fact, consistent with the principles of probability theory and Bayesian inference.

The second point considers recent findings by Nosofsky (2015) that a version of the exemplar model can account for existing inference patterns. We show that in discrete environments for certain values of the parameters of the exemplar model,

it makes the same predictions as the exiting models including the rational model. Therefore, Nosofsky's exemplar model can be viewed as a model that encompasses all possible strategies for feature-based inference with the existing models being the extreme cases. In continuous environments, however, one needs to make additional assumptions on the form of the response function for the predictions of the exemplar and the rational models to be equivalent.

Overall, currently available evidence suggests that the rational model makes predictions which are generally consistent with participants' inferences in discrete environments, both when conditional independence holds (Experiment 5 in Chapter 2) and when it does not hold (Murphy & Ross, 1994; Hayes & Newell, 2009; O. Griffiths et al., 2012, 2011). It also performs well in continuous environments with conditional independence (Experiments 1-4 in Chapter 2). There exists no evidence about how the rational model performs in continuous environments without conditional independence. It is important to measure the performance of the rational model in such environments because when features are continuous the predictions of different models are clearly distinct. We fill this gap and provide experimental evidence of the performance of the rational model in continuous environments. Consistent with our analysis of the existing findings, we find that the rational model performs better than the adaptation of the existing models to continuous environments.

In the following, we first describe the rational feature inference model in general terms. Then we focus on discrete environments. We show that in this case, the model reduces to the feature conjunction strategy discussed in the prior literature. We also review prior experiments and show that participants' inferences were consistent with the predictions of this model in many of these experiments. Next, we focus on continuous environments. We show how the existing model can be adapted to the continuous setting and report an experiment in which we measure the performance of the competing models. Finally, we discuss the relation of these models to Nosofsky's exemplar model and conclude.

3.2 Rational feature inference

3.2.1 Representing Mental Categories

Following the notation in Chapter 2, we model mental categories using probability distribution functions (*pdfs*) on the feature space.

3.2.2 Anderson's Model: Rational Inferences with Conditional Independence

Suppose that an individual observes that the first feature has value x and predicts the value of the second, unobserved feature. The model specifies her posterior distribution for the value of the second feature given her observation of the value of the first feature: $f(y \mid x)$ (see a detailed description of the Anderson's model

in Section 2.3.2 in Chapter 2). This model makes an important assumption: the *conditional independence assumption*. In other words, it assumes that the within-category feature correlation is zero. Expressed differently, it assumes that $f(y \mid xc) = f(y \mid c)$.

3.2.3 The rational model: feature inferences with or without conditional independence

In the generalized rational model, we specify the posterior on the second feature given the first feature by using the law of total probabilities:

$$f(y \mid x) = \sum_{c \in C} p(c \mid x) f(y \mid cx), \tag{3.1}$$

where $p(c \mid x)$ is the subjective probability that the object comes from category c given the observed feature value x on the first dimension (see eq. 2.6). $f(y \mid cx)$ is the marginal distribution of the value of the second feature, conditional on the fact that the object is in the category c and that its first feature has value x.

The rational model is psychologically realistic to the extent the agent can compute the components of the right hand side of eq. 3.1 on the basis of her mental representations and, moreover, do that in a way that is consistent with the rules of probability calculus. Currently available evidence reviewed in the next sections and experimental results reported below suggest that this is the case.

It is important to note that the rational model is a *computational model*: it specifies the computations the mind performs, but it does not specify a cognitive algorithm that could produce these computations (see Marr (1982) and T. L. Griffiths and Tenenbaum (2009) for enlightening discussions about computational versus algorithmic levels of explanations). By contrast, exemplar models are algorithmic models that specify how a cognitive system can compute the posterior probability (or an approximation of it) based on the data it has observed and memorized. We will show below that under some conditions, Nosofsky's model (2015) can be seen as an algorithmic model consistent with the rational computational model we just presented.

3.3 Rational Feature Inferences in Discrete Environments

3.3.1 Representing Categories

Consider a setting where objects have two discrete-valued features X and Y. For example, objects under consideration are colored shapes. They are organized into sets of objects - mental categories - each represented by their two features (x_i, y_i). Now suppose that when the individual encounters an object, she observes the value on the first feature $X = x$, say the shape, and wants to predict the value of Y, its color.

3.3.2 Rational Feature Inference

Equation 2.2 suggests a rational way to compute this. Adapting the notation to reflect the discrete nature of the environment, we define $P(c \mid x)$ as the proportion of objects that belong to c out of all the objects such that $X = x$, and $P(y \mid cx)$ as the proportion of objects with $Y = y$ out of the objects that both are in c and have $X = x$. In the equation, $P(y \mid cx)$ corresponds to the association between X and Y *within* category c. The weight of the prediction conditional on category c is given by application of Bayes' theorem

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}, \tag{3.2}$$

where $P(x \mid c)$ is the proportion of objects that have $X = x$ in category c. $P(c)$ is the prior about the category c.

Another approach is to use a different formula:

$$P(y \mid x) = \frac{P(xy)}{P(x)} = \frac{P(xy)}{\sum_y P(xy)}, \tag{3.3}$$

where $P(x)$ is the proportion of objects she has experienced with $X = x$ and $P(yx)$ is the proportion of objects she has experienced with $X = x$ and $Y = y$. In other words, the prediction is based on the feature correlations computed over all the data. This second approach, first introduced by Hayes et al. (2007), ignores categories altogether and has been called the 'feature conjunction' strategy in prior literature.

The two approaches are strictly equivalent as per the rules of probability calculus. This implies that in discrete environments, the rational model is equivalent to the 'feature conjunction' inferential strategy.

3.3.3 Existing Models

Single Category - Independent Features Another widespread approach proposes that people rely only on most likely category. In the first studies on the topic, (Murphy & Ross, 1994) introduced a model that ignores the within-category correlation (just like Anderson's model) and only uses the information from the most likely category. That is, first the individual determines which category is the most likely one (referred to as the 'target' category in the literature). Let us denote the 'target' category by c^* and let it be such that,

$$P(c^* \mid x) = \max_{c \in C} P(c \mid x). \tag{3.4}$$

then in the second stage, the individual computes the probability that the second feature is equal to y given that the item is a member of category c^*:

$$P(Y = y \mid x) = P_{c^*}(Y = y) \tag{3.5}$$

This approach is similar to Anderson's model but puts all the weight on the most likely category.

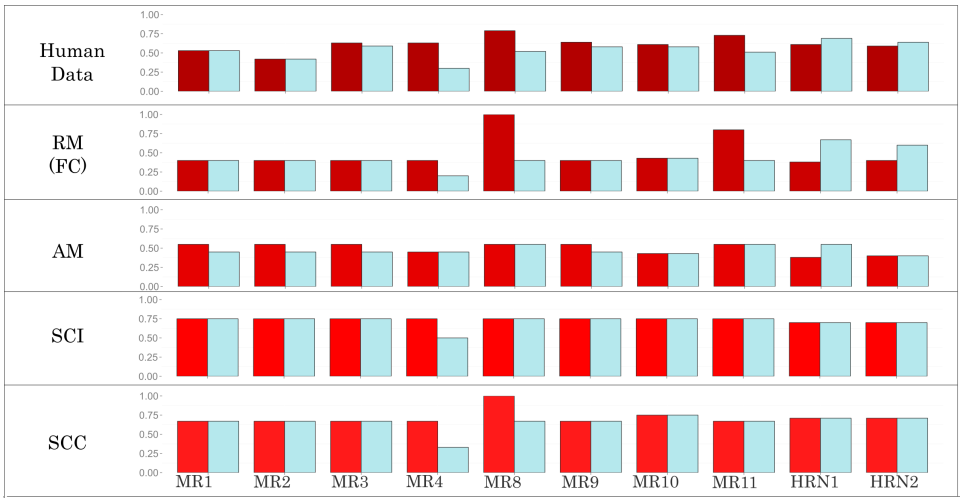


Figure 3.1: Human data and model predictions for experiments about feature inference with uncertain categorization: Probability Estimates. 10 Studies were used in the analysis. *MR* indicates the studies reported in Murphy & Ross, 1994; *HRN* indicated the studies reported in Hayes, Ruthven & Newell, 2007. The top panel depicts the average (across participants) probability estimates in two conditions (red and blue). Lower panels show the predictions of the models. *RM*: Rational Model; *FC*: Feature Conjunction; *AM*: Anderson's Model; *SCI*: Single Category Independent Features; *SCC*: Single Category Correlated Features.

Single Category - Correlated Features Another version of the single-category model relaxes the assumption about conditional independence and is sensitive to within-category correlations. This approach is similar to our rational model, but with all the weight put on the most likely category. In this case,

$$P(y \mid x) = P(y \mid c^* x). \tag{3.6}$$

3.3.4 Reinterpretation of Existing Findings

Existing Empirical Paradigm

The paradigm that was used in the majority of studies on feature-based inference with uncertain categorization was based on sets of objects with discrete-valued features (see Section 2.2 for a description of the paradigm). In all studies, participants were asked the following inference question: What is the most likely value of feature 2 given that feature 1 has value 'x'? In some studies, participants were also asked to report the probability of the unobserved feature given the value of the observed one: What is the probability that feature 2 has value 'y'? Due to this difference across studies, we provide two distinct analyses. First, we discuss

the results of the studies where probability predictions were available. Note, that this measure allows for a better comparison between the models as it does not require any additional assumptions about the response function. Second, we turn to the studies where only choices were recorded and assess the performance of the models under two different assumptions about the response function.

Out of all existing experiments, the analysis includes only those that fulfill two criteria. First, the exact display used in the experiment was available in the paper.[1] Second, we excluded studies for which the report of the results did not contain the actual choice proportions or the probability judgments (where applicable).[2]

Model Comparisons

Probability Estimates We identified ten experiments in which the participants were asked to provide subjective probability estimates. These studies are experiments 1, 2, 3, 4, 8, 9, 10 and 11 in Murphy and Ross (1994) and experiments 1 and 2 in Hayes et al. (2007). In each experiment, there were two conditions which corresponded to a specific question (e.g. 'what is the shape of an orange figure?'). Participants first reported the feature they believed to be the most likely and then the probability estimates for the feature they selected. An important feature of the paradigm is that the results were based on the comparison between two conditions. The upper panel of Figure 3.1 reports the average (across participants) estimates of the probability for all the mentioned experiments (in red for one condition and blue for the other condition). For example, the first pair of bars indicates that in Exp. 1 in Murphy and Ross (1994) in both conditions the average probability estimate is 53%.

The conditions were designed such that models made different predictions whether the difference between the conditions existed. For example, AM predicts a significant difference in Exp.1 Murphy and Ross (1994) whereas all other models predict no difference. The second panel reports the predictions of the Rational Model (RM). As we noted above they are equivalent to the predictions of the Feature Conjunction model (FC). The lower panels report the predictions of Anderson's Model (AM), Single Category Independent Features model (SCI), and Single Category Correlation Features model (SCC).

The rational model is the only model that provides a correct prediction in all ten experiments. In particular, no other model predicts the significant difference in the probability judgments in Exp.11 in Murphy and Ross (1994) and Exp.2 in Hayes et al. (2007). SCC correctly predicts the difference in only seven experiments. Good performance of both RM and SCC suggests that people are sensitive to feature correlation both between and within the categories. Analysis of the performance of AM and SCI confirms the poor fit of models that rely on the conditional inde-

[1] We did not include studies for which we could not reliably determine the structure of stimuli because in such cases we cannot estimate the precise predictions of the models.

[2] In particular, we did not include studies that reported the results using graphs (e.g. O. Griffiths et al. (2012)) without mentioning the exact numbers in the text or the graph.

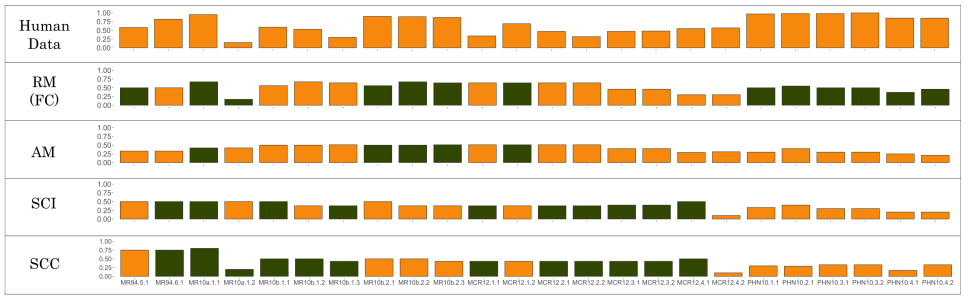


Figure 3.2: Human data and model predictions for experiments about feature inference with uncertain categorization: Choice data - Maximum Probablity Decision Rule. 24 Studies were considered in the analysis. Each bar corresponds to one study. The top panel depicts the choice proportions of the participants in each study. Lower panels show the predictions of the models. *RM*: Rational Model; *FC*: Feature Conjunction; *AM*: Anderson's Model; *SCI*: Single Category Independent Features; *SCC*: Single Category Correlated Features. If the predictions of the model are consistent with the choice data, it shaded dark green.

pendence assumption. AM and SCI predict correctly in two and six experiments respectively. Yet, in none of them, the models are the unique ones making a correct prediction.

Choice data The remaining studies included in the analysis only collected choice proportions: experiments 5 and 6 in Murphy and Ross (1994), experiment 1 in Murphy and Ross (2010a), experiments 1 and 2 in Murphy and Ross (2010b), experiments 1 to 4 in Murphy et al. (2012), and experiments 1 to 4 in Papadopoulos et al. (2011). In each experiment, we treated different conditions as different studies and compared how the models predict the choice proportions for each study. Overall, we analyze 24 studies.[3]

Because all the models are defined in terms of the probability distribution over the unobserved feature value, we needed to make additional assumptions about how this probability translates into choices. One simple version of such response function would assume that the feature with the highest probability is chosen. Then a model is said to make predictions that are consistent with the choice data if the

[3]We code each study the following way: first letters of the authors last names, year of the publication, number of an experiment, index of the condition in that experiment. For example, "MCR12.1.1" refers to condition 1 in Experiment 1 in Murphy, Chen, and Ross (2012). From left to right the bars on Figure 3.2 and Figure 3.3 correspond to experiments 5 and 6 Murphy and Ross (1994), conditions 1 and 2 in experiment 1 in Murphy and Ross (2010a), conditions 1,2,3 in experiments 1 and 2 in Murphy and Ross (2010b), conditions 1 and 2 in experiments 1 to 4 in Murphy et al. (2012), experiments 1 and 2 in Papadopoulos et al. (2011), conditions 1 and 2 in experiments 3 and 4 in Papadopoulos et al. (2011).

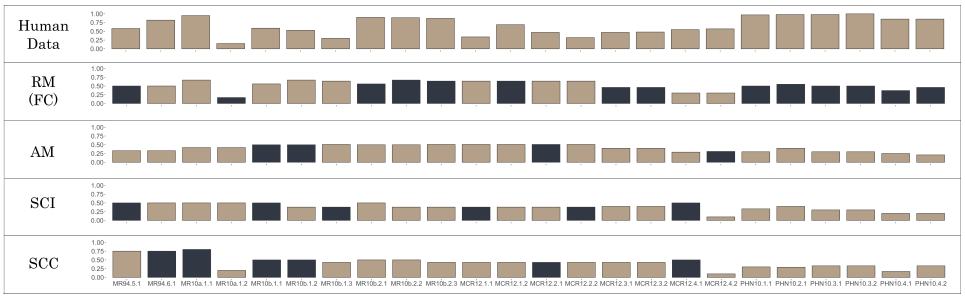


Figure 3.3: Human data and model predictions for experiments about feature inference with uncertain categorization: Choice data - Probablity Matching Decision Rule. 24 Studies were considered in the analysis. The top panel depicts the choice proportions of the participants in each study. Lower panels show the predictions of the models. *RM*: Rational Model; *FC*: Feature Conjunction; *AM*: Anderson's Model; *SCI*: Single Category Independent Features; *SCC*: Single Category Correlated Features. If the predictions of the model are consistent with the choice data, it shaded dark blue.

feature with the highest probability according to this model is chosen by the majority of the participants. Note that it is possible that more than one model makes the correct prediction. Consider a case, where according to one model the probability that a circle is red is 0.7 whereas according to another model this probability is 0.6. Now suppose that 60% of the participants chose red. Predictions of both models are consistent with the choice data. We refer to this as 'maximum probability' decision rule.

Under this assumption, the rational model's prediction is consistent with the choices of the majority of participants in 13 studies out of 24 (see Figure 3.2). Out of those, in seven studies RM is the only model that predicts correctly the choices of the majority of the participants. SCC does not perform as well. Although its predictions are consistent with participants' choices in 12 studies, there is only one study for which SCC the only best fitting model. Models that assume conditional independence again do not provide a good fit to the data. AM makes predictions that are consistent with the choice data in only 5 studies. SCI performs better and makes predictions that are consistent with the choices in 10 studies, in all of which, however, the choice patterns are also consistent with predictions of SCC.

This is not the only possible form of the response function. Another possibility is that people choose the feature according to its probability. We call this form of response function 'probability matching' decision rule. For example, the probability that a circle is red is 0.6 then the participant would choose red only 60% of the time. This assumption means that the prediction of the model in terms of probability is about the proportion of the participants that would choose this feature. Therefore, for each study, we calculated which model's probability prediction is the closest to the proportion of participants that chose that feature.

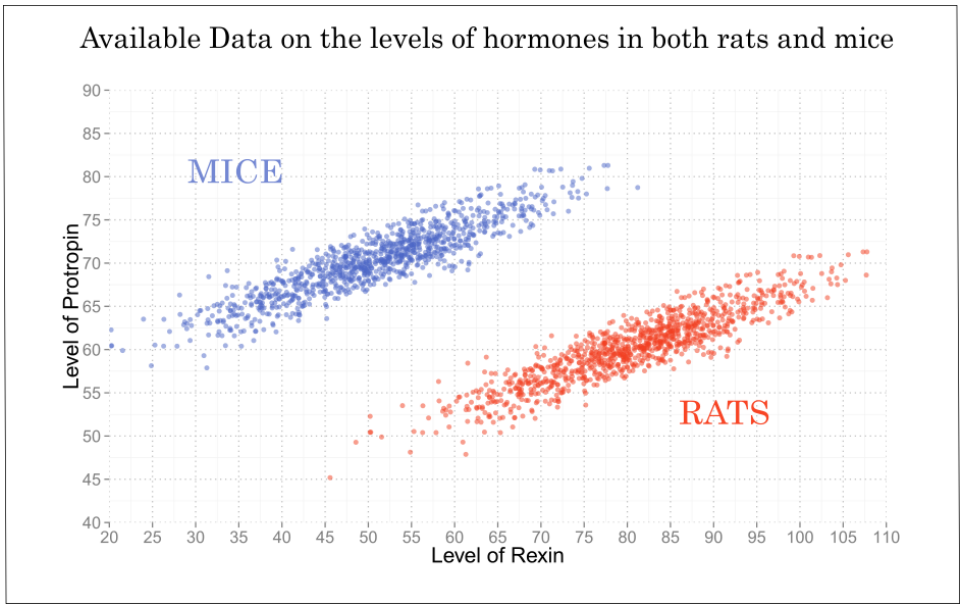


Figure 3.4: Categories used in the experiment. Participants were shown the level of 'Rexin' (x-axis) and were asked to predict the level of 'Protropin' (y-axis). The categories are 'rat' (R) and 'mouse' (M). $\mu_{xR} = 80, \mu_{yR} = 60, \mu_{xM} = 50, \mu_{yM} = 70, \sigma_{xR} = \sigma_{xM} = 10, \sigma_{yR} = \sigma_{yM} = 4, \rho_R = \rho_M = 0.9$.

Under this assumption, the rational model the best fitting model in 14 studies whereas AM, SCI, and SCC are the best in 4, 6, and 6 studies respectively (see Figure 3.3). Note that under 'probability matching' decision rule, the best model criterion is stricter than under the 'maximum probability' rule. In the former case, there are only 4 studies where more than one model makes predictions consistent with the data. In the latter case, there are 16 such studies.

Summary Reinterpretation of the existing studies shows that the predictions of the rational model are consistent with much of the findings in both the probability estimates and the choice proportions. Furthermore, the good performance of the models that rely on feature correlation compared to others that assume conditional independence suggests that people do not assume the absence of feature correlation rather they are quite sensitive to it.

3.4 Rational Feature Inferences in Continuous Environments

3.4.1 Representing Mental Categories

In continuous environments, instead of a collection of objects, we define a category as a probability distribution function (*pdf*) over the feature space (Ashby & Alfonso-Reese, 1995). For simplicity, in what follows we assume there are two relevant categories ($\mathcal{C} = \{1,2\}$) each represented by a bi-variate normal distribution:

$$\begin{pmatrix} X_c \\ Y_c \end{pmatrix} \sim N\Big(\begin{pmatrix} \mu_{xc} \\ \mu_{yc} \end{pmatrix}; \begin{pmatrix} \sigma^2_{xc} & \rho_c \sigma_{xc} \sigma_{yc} \\ \rho_c \sigma_{xc} \sigma_{yc} & \sigma^2_{yc} \end{pmatrix} \Big), \tag{3.7}$$

where μ_{xc} and μ_{yc} are the category means on the two features, σ_{xc} and σ_{yc} are the standard deviations on the two features and ρ_c is the within-category correlation for category c (see Figure 3.4 for an example). This definition is the same as the one used in the previous chapter (see Section 2.3.1) but it allows for within-category correlations.

3.4.2 Rational Feature Inference

The Rational Model specifies the posterior on the value of the second feature conditional on the value of the first feature by integrating the information over both categories. We adapt eq. 2.1 to a continuous setting with two categories and normally distributed *pdfs*:

$$\begin{aligned} f(y \mid x) = P(c_1 \mid x) f_{\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}} \rho_{c_1} (x - \mu_{xc_1}), \sigma_{y1}}(y) \\ + P(c_2 \mid x) f_{\mu_{yc_2} + \frac{\sigma_{yc_2}}{\sigma_{xc_2}} \rho_{c_2} (x - \mu_{xc_2}), \sigma_{y2}}(y), \end{aligned} \tag{3.8}$$

where $f_{\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}} \rho_{c_1} (x - \mu_{xc_1}), \sigma_{y1}}(y)$ defines a normal density with a mean $\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}} \rho_{c_1} (x - \mu_{xc_1})$ and standard deviation σ_{y1}. Furthermore, the weights of the categories are defined by $P(c_1 \mid x) = 1 - P(c_2 \mid x)$ which is computed by applying Bayes' rule:

$$P(c_1 \mid x) = \frac{1}{1 + e^{ax^2 - bx + c}}, \tag{3.9}$$

with

$$\begin{aligned} a &= \frac{\sigma^2_{x2} - \sigma^2_{x1}}{2\sigma^2_{x2}\sigma^2_{x1}}, \\ b &= \frac{\sigma^2_{x2}\mu_{x1} - \sigma^2_{x1}\mu_{x2}}{\sigma^2_{x2}\sigma^2_{x1}}, \\ c &= \frac{\sigma^2_{x2}\mu^2_{x1} - \sigma^2_{x1}\mu^2_{x2}}{2\sigma^2_{x2}\sigma^2_{x1}} + \log \frac{\sigma_{x2}}{\sigma_{x1}} + \log \frac{P(c_2)}{P(c_1)}. \end{aligned}$$

Here we assume that the priors about the categories are equal: $P(c_1) = P(c_2) = 0.5$. This model predicts that the inference about the position of the object on the second feature is influenced both by the category level information (likelihoods that the object came from one category or another) and the internal structure of the categories (the within-category correlation between X and Y). We will refer to this model as the rational model (RM).

Relation to Anderson's Model

Here we use the adaptation of Anderson's model to continuous environments proposed in the previous chapter (see Section 2.3.2). Under the conditional independence assumption, the rational model is the same as Anderson's model (AM). In terms of our model, it amounts to assuming $\rho_{c1} = \rho_{c2} = 0$.

3.4.3 Other models

Single Category - Independent Features (SCI)

Here we use the same formulation of the model that relies only on the most likely category as described in Chapter 2 Section 2.3.3.

Single Category - Correlated Features (SCC)

The version of the single category model that is sensitive to the within-category correlation is characterized by $f(y \mid x) = f_{c^*}(y \mid x)$, where

$$f_{c^*} = f_{\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}} \rho_{c_1}(x - \mu_{xc_1}), \sigma_{y1}} \tag{3.10}$$

if the target category is category 1, and

$$f_{c^*} = f_{\mu_{yc_2} + \frac{\sigma_{yc_2}}{\sigma_{xc_2}} \rho_{c_2}(x - \mu_{xc_2}), \sigma_{y2}}, \tag{3.11}$$

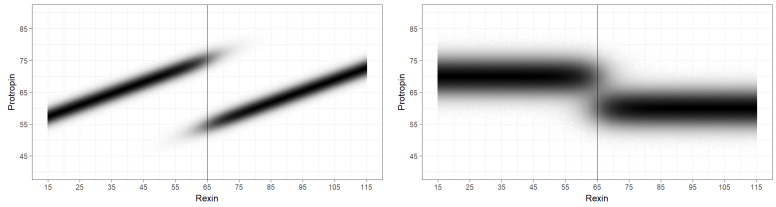
otherwise. This model is then similar to the Rational Model but puts all the weight on the 'target' category.

Ignoring the Categories: Linear Model (LM)

Here we use the same linear model described in Chapter 2 Section 2.3.3.

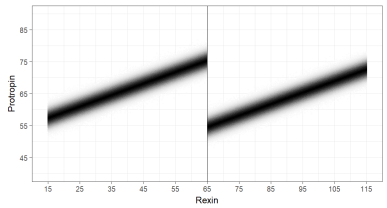
3.4.4 Decision Rule

All the competing models are described in terms of the posterior distributions of the unobserved feature Y given the value of X. In order to make empirical predictions about human inferences, we need to make an additional assumption about how the posterior is translated into the human responses. In the analysis of the

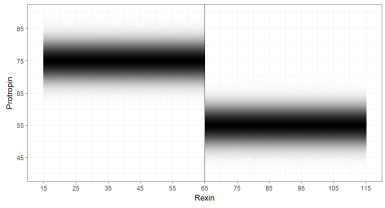


(a) RM: Rational Model

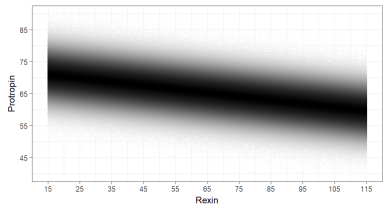
(b) AM: Anderson's Model.



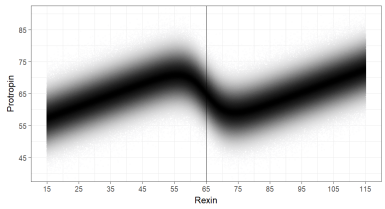
(c) SCI: Single Category Correlated Features



(d) SCI: Single Category Independent Features



(e) LM: Linear Model



(f) RMD: Rational Model Deterministic

Figure 3.5: Posteriors ($f(y \mid x)$) on the second feature (Rexin) given the value of the first feature (Protropin) of the competing models. The darker areas correspond to higher values of the posterior and lighter areas correspond to lower values. The vertical line at $x = 65$ represents the 'boundary', i.e., the value of x for which the two categories are equally likely $P(\text{'rat'} \mid x = 65) = P(\text{'mouse'} \mid x = 65)$.

models' performance, we assume that the inferred value of the unobserved feature is a random draw from the posterior distribution - this is a 'probability matching' decision rule discussed in the previous section.

The rational model makes an important assumption that the individual integrates across the available categories. With the 'probability matching' rule, we assume that the integration happens at the stage of computation of the posterior but not on the level of the decision rule. Drawing on the discussion in Section 2.10.1 of the previous chapter, we also analyze the performance of the model that assumes response function to be the expected value of the posterior $E[Y \mid x]$ computed according to the rational model. We will refer to this model as 'RM Deterministic' (RMD). For the setup with two categories with normally distributed *pdfs*, the ex-

pected value is:

$$E[Y \mid x] = P(c_1 \mid x)\mu_{yc_1} + \frac{\sigma_{yc_1}}{\sigma_{xc_1}}\rho_{c_1}(x-\mu_{xc_1}) + P(c_2 \mid x)\mu_{yc_2} + \frac{\sigma_{yc_2}}{\sigma_{xc_2}}\rho_{c_2}(x-\mu_{xc_2}). \tag{3.12}$$

We specify this model using a normally distributed error term to account for the possibility of human errors. Then the posterior for RMD is:

$$f(y \mid x) = f_{E[Y|x],\sigma_{RMD}}(y) \tag{3.13}$$

where $f_{E[Y|x],\sigma_{RMD}}$ is a normally distributed variable with mean $E[Y \mid x]$ defined in eq. 3.12 and standard deviation σ_{RMD}.

3.5 Experiment

Similarly to the experiments reported in Chapter 2, participants were making feature-based inference with uncertain categorization using artificial categories.

3.5.1 Design

The experiment replicates the paradigm developed in Experiment 1 in the Chapter 2 with one change: within-category feature correlation is positive in both categories (for details about the design see section 2.4.1). 29 participants were recruited via Amazon Mechanical Turk.

3.5.2 Model Predictions

Figure 3.5 illustrates the posterior distributions $f(y \mid x)$ implied by the five competing models. In the first row, predictions of RM and AM are depicted. The posteriors of RM and AM are based on eq. 3.8 and eq. 2.7 with category weights defined by eq. 3.9. In the second row, the models based on single category rely on the same posteriors but with the all-or-nothing category weights. Parameters that were used in the models were also used to generate the category data (see legend of Figure 3.4). In the last row, the posterior for LM is based on eq. 2.9. The parameters are the coefficients of the linear regression model based on all the data on Figure 3.4. Based on the data used in the experiment, the parameters are $a_0 = 72.7$, $a_1 = -0.15$, and $\sigma_l = 6.1$. The posterior for RMD is based on eq: 3.13. Parameter σ_{RMD} was obtained by maximum likelihood estimation of the posterior with just σ_{RMD} as a free parameter and the true values for the other parameters (see legend of Figure 3.4), based on all the exemplars depicted on Figure 3.4 (irrespective of their categories).

Table 3.1: Percentage of participants whose feature predictions were best fit by each of the candidate models.

Model	True	Est.	True	Est.
RM: Rational	15(53%)	18(62%)	15(52%)	17(59%)
AM: Anderson	1(3%)	0	0	0
SCC: Single Cat. Corr. Features	5(17%)	8(28%)	5(17%)	7(24%)
SCI: Single Cat. Indep. Features	1(3%)	0	0	0
LM: Linear	7(24%)	3(10%)	4(14%)	3(10%)
RMD: Rational Deterministic	-	-	5(17%)	2(7%)
# participants	29		29	

The crucial area is around the point where the categories are equally likely ($P(c_1 \mid x) = 0.5$ or $x = 65$). Under the 'probability matching' decision rule, the rational model and Anderson's model predict that values from both categories are likely whereas according to the single category models values only from the most likely category are likely. The difference within the pairs of models is whether the prediction is also sensitive to the within-category correlation. Consider Rexin level of 60, the Rational Model predicts that values around *both* 53 and 73 are likely (consistent with the within-category correlations). Anderson's model, in this case, predicts that values around both means are likely (60 and 70 for 'Rats' and 'Mice' respectively). For this value of Rexin, the most likely category is 'Mice', therefore according to the single category models only values from the 'Mice' category are likely. In particular, SCI predicts that levels of Protropin around 70 are likely and SCC predicts that values around 73 are likely. In contrast, RMD predicts values that are between the means of the categories. For Rexin level of 60, this model predicts values close to 69.

3.5.3 Results

Parameter-Free Model Comparison Here we assume that the parameter values used for all models apart from LM and RMD are the ones used to generate the category data. For LM and RMD the parameters were estimated as described in the previous section. For each model, we computed log-likelihood in the participant-by-participant basis. First, consider a comparison without RMD. The rational model provides the best fit to the majority of participants (53%). The single category model with feature correlation fits 17% of the participants. The performance of RM does not change when RMD is added to the comparison (see Table 3.1). Figure 3.6 shows inferences of each participant and the corresponding log-likelihoods

Table 3.2: Estimated model parameters. Parameters were estimated separately for each participant. The values are the mean estimated parameters across participants for whom that model is the best. RM: Rational model, SCC: the single category correlated features model; LM: linear model. AM (Anderson's model) and SCI (the single category independent features model) fitted 0 participants.

Parameter	$\mu_{x,R}$	$\mu_{y,R}$	$\mu_{x,M}$	$\mu_{y,M}$	σ_x	σ_y	ρ	a_0	a_1	σ_l
True Value	**80**	**60**	**50**	**70**	**10**	**4**	**0.9**	**72.7**	**-0.15**	**6.1**
Model										
RM	81.9	60.8	42.9	67.1	8.7	3.3	0.77			
AM	-	-	-	-	-	-	-	-	-	-
SCC	80.5	60.5	50.3	69.8	9.6	3.7	0.88			
SCI	-	-	-	-	-	-	-	-	-	-
LM								53.9	0.13	3.34

of all models.

Comparison of Models with Parameters Estimated Participant-by-Participant Following the same procedure as in the analysis of the results of four experiments in Chapter 2 Section 2.7.2, we estimated the parameters of all the models participant-by-participant using maximum likelihood.[4] This allows the comparison to account for the individual interpretation of the category information.

The average estimates of the parameters are quite close to the real parameters which indicates that people understood quite well the structure of the categories (see Table 3.2 for average estimates of the parameters across participants for whom that model is the best). When compared based on BIC criterion, the pattern of results is similar to the parameter-free comparison. Consider the first comparison without RMD. RM provides the best fit for the majority of participants (43%) and SCC provides the best for 37% of the participants. The models that rely on conditional independence provide the best fit for none of the participants. This pattern does not change significantly when RMD is included in the comparison.

Analyses of the 'switching' behavior of the participants Mirroring the analysis of the experiments in Section 2.7.2, we also report the 'switching' behavior of the participants. There are 11 participants that make exactly one switch. This is

[4]We used box-constrained optimization for all models apart from LM. The lower bound for all parameters was 0 and correlation parameter is also bounded by 1 for RM and SCC. RMD was optimized over all category parameters and σ_{RMD}. LM was optimized without constraints on its three parameters.

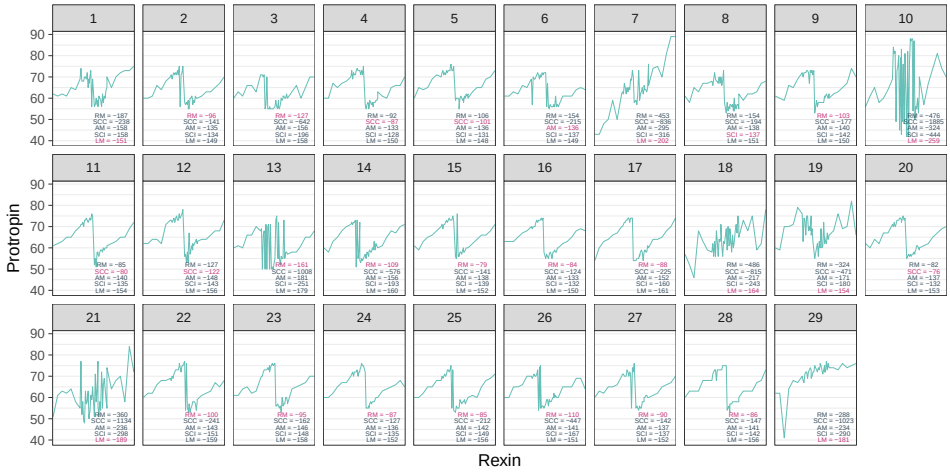


Figure 3.6: Inference of the participants in the experiment. The levels of likelihood in the parameter-free case of each model are shown on the graph. RM: The Rational Model, SCC: the single category correlated feature model; AM: Anderson's Model; SCI: the single category independent feature model; LM: linear model. The best fitting model is shaded in red.

similar to the number of participants best fitted by SCC (based on participant-by-participants estimated parameters - see Table 3.1).

3.5.4 Discussion

Model comparisons suggest that the rational model provides an appropriate characterization of the behavior of a large proportion of the participants when compared to other models. This implies that most participants consider the two candidate categories when making predictions about the unobserved feature. At the time, we find a similar extent of heterogeneity among the participants as in studies reported in the Chapter 2 and in existing studies on feature-based inference with uncertain categorization. Importantly, our results provide further evidence that the conditional independence assumption is unrealistic. Because both AM and SCI provide a very poor fit to the participants' inferences, it is clear that participants are sensitive to the within-category correlations when making inferences.

Furthermore, comparison with other models yields that RMD does not provide a good fit to the participants' data. For both sets of analysis, the number of participants best fit by RMD is considerably low than for the Rational Model. This provides evidence that participants integrated the information about the categories on the level of the posterior rather than the decision rule. This result replicates findings reported in Chapter 2 for environments with condition independence.

3.6 Relation to Nosofky's Exemplar Model

Nosofsky (2015) recently proposed an exemplar model that makes inferences consistent with much of the existing data on predictions with uncertain categorization. In this section, we first describe the model, then characterize its the relationship to the existing models both in discrete and continuous environments.

Like the models discussed above, the exemplar model provides an estimate of $P(y \mid x)$. According to Nosofsky, "the key idea is that when an observer makes an inference that an object belongs to a category, the inferred category label becomes a new feature of that object. The observer then assesses the similarity of the object/category-label pair to all other object-label pairs in the display, using the similarity rules formalized in the exemplar models of Medin and Schaffer (1978) and Nosofsky (1984). The probability that the subject infers that the newly queried feature is Feature X is then based on the summed similarity of the object-label pair to all object-label exemplars that contain Feature X (Medin & Schaffer, 1978; Nosofsky, 1984)."

The model assumes that the decision maker identifies the most likely category (the 'target' category) based on the observed feature x. In computing an estimate of $P(y \mid x)$, the model is sensitive to both the exemplars for which the first feature has value x and ones that belong to the target category.

Let $D_2(y)$ denote the set of exemplars with value y on the second dimension and let $\eta_{\mathbf{AB}}$ denote the similarity between any two exemplars $\mathbf{A}$ and $\mathbf{B}$. Let c^* refer to the target category (see eq. 3.4). Finally, we denote by $(x?c^*)$ an exemplar whose first feature is known to have value x and for which the most likely category is c^*, but for which the second feature is not observed.

According to the model:

$$\hat{P}(y \mid x) = \frac{\sum_{\mathbf{j} \in D_2(y)} \eta_{(x?c^*),\mathbf{j}}}{\sum_{\mathbf{j}} \eta_{(x?c^*),\mathbf{j}}}. \tag{3.14}$$

Here the numerator is the sum of similarities of an object $(x?c^*)$ to all objects for which the second feature has value y. In turn, the denominator is the sum of similarities of the same object $(x?c^*)$ to all exemplars regardless of the value of the second feature.

The similarity is computed from the interdimensional multiplicative rule of the exemplar model:

$$\eta_{\mathbf{AB}} = \eta_{x_A,x_B} \cdot \cdot \eta_{c_A,c_B},$$

The similarity depends on two parameters S and L. The first parameter S characterizes the weight of exemplars that do not have the observed feature. In particular, $\eta_{x_A,x_B} = 1$ if $x_A = x_B$, $\eta_{x_A,x_B} = S$ if $x_A \neq x_B$. This implies that when $S = 0$, only the exemplars that have the observed features are considered in the computation of the probability of the unobserved feature. When $S = 1$, all exemplars are considered in the inference of the unobserved feature.

The second parameter, L, regulates the sensitivity to the target category: $\eta_{c_A^*,c_B^*} = 1$ if $c_A^* = c_B^*$, and $\eta_{c_A^*,c_B^*} = L$ if $c_A^* \neq c_B^*$. Therefore, when $L = 0$ exemplars outside the target category are ignored. When $L = 1$, exemplars outside the target category are given the same weight as exemplars in the target category.

3.6.1 Discrete Environments

First, we discuss how the exemplar model relates to the existing models in discrete environments. Here, we show that for some combinations of parameters values (of S and L) the exemplar model makes the same predictions as several of the models discussed. Just as in the section on rational feature inference in discrete environments, the probabilities of different features and categories are expressed in empirical proportions. Let N be the total number of exemplars. The number of exemplars with value x on the first feature is thus $P(x)N$.

Next, we write the predicted conditional probability $\hat{P}(y \mid x)$ in terms of these frequencies. We start with the numerator:

$$\sum_{\mathbf{j} \in D_2(y)} \eta_{(x?c^*),\mathbf{j}} = P(x,y,c*)N + P(\neg x,y,c*)SN + L \sum_{c \neq c*} \big[P(x,y,c)N + P(\neg x,y,c)SN\big], \quad (3.15)$$

where $\neg x$ denotes any value of the first feature different from x. Similarly, we write the denominator as:

$$\sum_{\mathbf{j}} \eta_{(x?c^*),\mathbf{j}} = P(x,c*)N + P(\neg x,c*)SN + L \sum_{c \neq c*} \big[P(x,c)N + P(\neg x,c)SN\big]. \quad (3.16)$$

Now we will consider specific parameter values and show for which parameter values the prediction of the exemplar model corresponds to different existing models.

Case 1 First, suppose that $L = 1$. This implies the same weight is given to all exemplars or inferences should be insensitive to categorical boundaries.
The numerator becomes:

$$\begin{aligned}\sum_{\mathbf{j} \in D_2(y)} \eta_{(x?c^*),\mathbf{j}} &= N \sum_c \big[P(x,y,c) + P(\neg x,y,c)S\big], \\ &= N\big[P(x,y) + P(\neg x,y)S\big].\end{aligned}$$

The denominator becomes:

$$\begin{aligned}\sum_{\mathbf{j}} \eta_{(x?c^*),\mathbf{j}} &= N \sum_c \big[P(x,c) + P(\neg x,c)S\big], \\ &= N\big[P(x) + P(\neg x)S\big].\end{aligned}$$

Then,

$$\hat{P}(y \mid x) = \frac{P(x,y) + P(\neg x, y)S}{P(x) + P(\neg x)S}. \tag{3.17}$$

Case 1.1. Consider the model's prediction when $S = 0$. Then,

$$\hat{P}(y \mid x) = \frac{P(x,y)}{P(x)} = P(y \mid x). \tag{3.18}$$

In this case, the model's predicted conditional probability is exactly the empirical conditional probability - this is also the prediction of *RM* in discrete environments. This is consistent with intuition because $L = 1$ and $S = 0$ implies that all exemplars that have x are considered regardless of their category.

Case 1.2. Now consider instead $S = 1$. Then,

$$\hat{P}(y \mid x) = P(y). \tag{3.19}$$

The model's predicted conditional probability is exactly the prior on the unobserved feature. Here, the probability estimation is again insensitive to the category boundaries (as implied by $L = 1$), but is also insensitive to the feature correlation (as implied by $S = 1$).

Case 2 Now consider cases where $L = 0$. In this case, the weight of 0 is given to exemplars outside of the target category. The numerator becomes:

$$\sum_{\mathbf{j} \in D_2(y)} \eta_{(x?c^*),\mathbf{j}} = P(x,y,c*)N + P(\neg x,y,c*)SN.$$

The denominator becomes:

$$\sum_{\mathbf{j}} \eta_{(x?c^*),\mathbf{j}} = P(x,c*)N + P(\neg x,c*)SN.$$

Then,

$$\hat{P}(y \mid x) = \frac{P(x,y,c*) + P(\neg x,y,c*)S}{P(x,c*) + P(\neg x,c*)S}. \tag{3.20}$$

Case 2.1. First, suppose $S = 0$. Then,

$$\hat{P}(y \mid x) = \frac{P(x,y,c*)}{P(x,c*)} = P(y \mid x,c*). \tag{3.21}$$

Here the model's prediction is equivalent to the prediction of the *SCC* model. Intuitively, $L = 0$ implies that the exemplars outside of the target category are ignored and $S = 0$ implies that only the exemplars that match the observed feature value are considered.

Case 2.2. Second, suppose $S = 1$. Then,

$$\hat{P}(y \mid x) = \frac{P(x,y,c*) + P(\neg x,y,c*)}{P(x,c*) + P(\neg x,c*)} = \frac{P(y,c*)}{P(c*)} = P(y \mid c*). \tag{3.22}$$

In this case, the model's prediction matches the prediction of SCI. Here, not only the exemplars outside of the target category are ignored (as implied by $L = 0$) but also the feature conjunctions (as implied by $S = 1$).

It is worth noting that under conditional independence the exemplar model is equivalent to SCI whenever $L = 0$ (for all S values). To see why, note that conditional independence implies that for all c, $P(xy \mid c) = P(x \mid c)P(y \mid c)$. We can thus rewrite eq. 3.20 as

$$\begin{aligned}&\hat{P}(y \mid x)\\ &= \frac{P(x \mid c*)P(y \mid c*)P(c*) + P(\neg x \mid c*)P(y \mid c*)P(c*)S}{P(x \mid c*)P(c*) + P(\neg x \mid c*)P(c*)S}\\ &= P(y \mid c*).\end{aligned} \tag{3.23}$$

Relation to Anderson's Model Interestingly, there does not exist any combination of parameter values such that the exemplar model is equivalent to Anderson's model. Intuitively, this is because the exemplar model is always sensitive to within-category feature correlations whereas Andersons' model assumes that this correlation is 0 even if it is not the case in the environment. To see this, note that according to the exemplar model $\hat{P}(y \mid x)$ can be written as a ratio. To compare the predictions of the two models, we also write the posterior implied by Anderson's model as a ratio. We get

$$\hat{P}(y \mid x) = \sum_{c \in \mathcal{C}} P(c \mid x)P(y \mid c) = \frac{\sum_{c \in \mathcal{C}} P(c)P(x \mid c)P(y \mid c)}{P(x)}. \tag{3.24}$$

We use a *Reductio ad absurdum*. Let us assume there exist parameter values (of L and S) such that the two models produce the same posterior. In this case, the two denominators have to be proportional to $P(x)$. According to equation 3.16, the denominator of the posterior implied by the exemplar model can only be proportional to $P(x)$ if $L = 1$ and $S = 0$. But in this case, the numerator of the posterior according to the exemplar model is $P(x,y)$ and the posterior according to this model is the same as that produced by RM: $P(y \mid x)$. This is clearly different from Anderson's model since the two strategies are the same only in the special case where the environment is characterized by conditional independence.

3.6.2 Continuous Environments

The relation between the exemplar model and the existing models in continuous environments is not as straightforward. For instance, Shi et al. (2010) showed that the exemplar model can be seen as the algorithm for performing Bayesian

inference. Importantly, for the equivalence to exist one needs to make a very strict assumption about the decision rule. In particular, they showed the equivalence between predictions of the exemplar and the Bayesian models when the decision rule is assumed to be the expected value of the posterior produced by the rational model. We referred to this model as the 'RM Deterministic' in the discussion of the experimental results. Note that our experimental results do not support such assumption about the decision rule. Rather, they are consistent with a 'probability matching' form of response. Whether an equivalence exists under the 'probability matching' decision rule is not clear and requires further investigation.

3.7 Conclusion

Our rational model implies that when the category of the item is uncertain, participants should give weight to the predictions implied by membership in the two candidate categories. This should be the case both under conditional independence or when there is within-category feature correlation. The analysis of the existing studies on the topic as well as our empirical results suggest that a majority of participants behaved according to this qualitative prediction. At the same time, there is a certain amount of heterogeneity in the choice of inferential strategy. Although the majority of the participants followed the rules of probability calculus, a significant minority relied only on the most likely category to make inferences. Determining when people use one strategy or another is an interesting avenue for future investigation.

Additionally, we provided further evidence that people are sensitive to within-category correlations when making inferences. One, however, might wonder if it is a function of the complexity of the task at hand. Whether such sensitivity persists when objects have more than two dimensions or organized into more than two or four categories remains to be explored.

Finally, our model is a computational model and, as such, it does not specify how people might perform the computations that lead to these predictions (Marr, 1982). The exemplar model that Nosofsky (2015) recently proposed is an algorithmic model that achieves such predictions. The computational analysis of this model in discrete environments illustrates that one can treat the exemplar model as a more general model that encompasses all the existing inferential strategies which in turn represent the extreme strategies on the spectrum. By contrast, in continuous environments, additional assumptions about the decision rule are needed to establish equivalence between the rational model and the exemplar model. Current work does not support the required equivalence assumption which implies that further work is needed to clarify the relationship between the 'computational' rational model and the 'algorithmic' exemplar model.

References

Alves, H., Koch, A., & Unkelbach, C. (2016). My friends are all alike the relation between liking and perceived similarity in person perception. *Journal of Experimental Social Psychology*, *62*, 103–117.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Anderson, J. R., & Fincham, J. M. (1996). Categorization and sensitivity to correlation. *Journal of experimental psychology: Learning, Memory, and Cognition*, *22*(2), 259.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.

Badea, C., Brauer, M., & Rubin, M. (2012). The effects of winning and losing on perceived group variability. *Journal of Experimental Social Psychology*, *48*(5), 1094–1099.

Beach, L. R., & Scopp, T. S. (1968). Intuitive statistical inferences about variances. *Organizational Behavior and Human Performance*, *3*(2), 109–123.

Boldry, J. G., & Gaertner, L. (2006). Separating status from power as an antecedent of intergroup perception. *Group processes & intergroup relations*, *9*(3), 377–400.

Boldry, J. G., Gaertner, L., & Quinn, J. (2007). Measuring the measures a meta-analytic investigation of the measures of outgroup homogeneity. *Group Processes & Intergroup Relations*, *10*(2), 157–178.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate. *Journal of social issues*, *55*(3), 429–444.

Brunswick, E. (1952). *The conceptual framework of psychology.* Chicago: The University of Chicago Press.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury Pacific Grove, CA.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014a). Decision making under uncertain categorization. *Frontiers in psychology*.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014b). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General*, *143*(1), 227.

Crawford, L. E., Huttenlocher, J., & Hedges, L. V. (2006). Within-category feature correlations and bayesian adjustment strategies. *Psychonomic Bulletin & Review*, *13*(2), 245–250.

Davis, J. A., Smith, T. W., & Marsden, P. V. (2016). *General social survey, 1993, 1998, 2000, 2002 with cultural, information security, and freedom modules [united states].* Inter-university Consortium for Political and Social Research (ICPSR). Retrieved from `https://doi.org/10.3886/ICPSR35536.v2` (ICPSR35536-v2)

Denrell, J. (2005). Why most people disapprove of me: experience sampling in impression formation. *Psychological review*, *112*(4), 951.

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological review*, *85*(5), 395.

Fazio, R. H., Eiser, J. R., & Shook, N. J. (2004). Attitude formation through exploration: valence asymmetries. *Journal of personality and social psychology*, *87*(3), 293.

Fiedler, K. (2000). Beware of samples! a cognitive-ecological sampling approach to judgment biases. *Psychological review*, *107*(4), 659.

Fiedler, K. (2012). Meta-cognitive myopia and the dilemmas of inductive-statistical inference. In *Psychology of learning and motivation-advances in research and theory* (Vol. 57, p. 1).

Fiedler, K., & Juslin, P. (2006a). *Information sampling and adaptive cognition*. Cambridge University Press.

Fiedler, K., & Juslin, P. (2006b). Taking the interface between mind and environment seriously. *Information sampling and adaptive cognition*, 3–29.

Fox, C. R., & Hadar, L. (2006). " decisions from experience"= sampling error+ prospect theory: Reconsidering hertwig, barron, weber & erev (2004). *Judgment and Decision Making*, *1*(2), 159.

Galesic, M., Olsson, H., & Rieskamp, J. (2012). Social sampling explains apparent biases in judgments of social environments. *Psychological Science*, 0956797612445313.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press.

Gigerenzer, G., Todd, P. M., & Group, A. R. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making*, *9*(1), 1.

Griffiths, O., Hayes, B. K., & Newell, B. R. (2012). Feature-based versus category-based induction with uncertain categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 576–595.

Griffiths, O., Hayes, B. K., Newell, B. R., & Papadopoulos, C. (2011). Where to look first for an explanation of induction with uncertain categories. *Psychonomic bulletin & review*, *18*(6), 1212–1221.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*(4), 661–716.

Hamilos, C. A., & Pitz, G. F. (1977). The encoding and recognition of probabilistic information in a decision task. *Organizational Behavior and Human Performance*, *20*(2), 184–202.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, *10*(3), 252–264.

Haslam, S. A., Oakes, P. J., Turner, J. C., & McGarty, C. (1995). Social categorization and group homogeneity: Changes in the perceived applicability of stereotype content as a function of comparative context and trait favourableness. *British Journal of Social Psychology*, *34*(2), 139–160.

Hayes, B. K., & Chen, T.-H. J. (2008). Clinical expertise and reasoning with uncertain categories. *Psychon Bull Rev*, *15*(5), 1002-7. doi: 10.3758/PBR.15.5 .1002

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition*, *37*(6), 730–743.

Hayes, B. K., Ruthven, C., & Newell, B. R. (2007). Inferring properties when categorization is uncertain: A feature-conjunction account. In *Proceedings of the 29th annual conference of the cognitive science society* (pp. 209–214).

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological science*, *15*(8), 534–539.

Hogarth, R. M. (1975). Cognitive processes and the assessment of subjective probability distributions. *Journal of the American statistical Association*, *70*(350), 271–289.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive psychology*, *24*(1), 1–55.

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological review*, *114*(3), 733.

Jones, E. E., Wood, G. C., & Quattrone, G. A. (1981). Perceived variability of personal characteristics in in-groups and out-groups: The role of knowledge and evaluation. *Personality and Social Psychology Bulletin*, *7*(3), 523–528.

Judd, C. M., & Park, B. (1988). Out-group homogeneity: Judgments of variability at the individual and group levels. *Journal of Personality and Social Psychology*, *54*(5), 778–788.

Judd, C. M., Ryan, C. S., & Park, B. (1991). Accuracy in the judgment of in-group and out-group variability. *Journal of personality and social psychology*, *61*(3), 366.

Juslin, P., Winman, A., & Hansson, P. (2007). The naive intuitive statistician: a naive sampling model of intuitive confidence intervals. *Psychological review*, *114*(3), 678.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition*, *56*(3), 263–269.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological review*, *107*(2), 397.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, *131*(2), 287.

Kemp, C., Shafto, P., & Tenenbaum, J. B. (2012). An integrated account of generalization across objects and features. *Cognitive Psychology*, *64*(1), 35–73.

Konovalova, E., & Le Mens, G. (2016). Predictions with Uncertain Categorization: A Rational Model. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th annual conference of the cognitive science society* (pp. 722–727). Austin, TX.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, *108*(3), 480.

Le Mens, G., & Denrell, J. (2011, April). Rational learning and information sampling: On the "naivety" assumption in sampling explanations of judgment biases. *Psychological Review*, *118*(2), 379–392.

Le Mens, G., Kareev, Y., & Avrahami, J. (2016). The evaluative advantage of novel alternatives: An information-sampling account. *Psychological science*, *27*(2), 161–168.

Linville, P. W., & Fischer, G. W. (1998). Group variability and covariation: Effects on intergroup judgment and behavior. *Intergroup cognition and intergroup behavior*, 123–150.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of personality and social psychology*, *57*(2), 165.

Linville, P. W., Fischer, G. W., & Yoon, C. (1996). Perceived covariation among the features of ingroup and outgroup members: The outgroup covariation effect. *Journal of Personality and Social Psychology*, *70*(3), 421.

Lorenzi-Cioldi, F. (1998). Group status and perceptions of homogeneity. *European review of social psychology*, *9*(1), 31–75.

Malt, B. C., Ross, B. H., & Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(3), 646–661.

Malt, B. C., & Smith, E. (1984). Correlated properties in natural categories. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 250–269.

March, J. G. (1996). Learning to be risk averse. *Psychological review*, *103*(2), 309.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. NewYork: Freeman.

Marsden, P. V. (1987). Core discussion networks of americans. *American sociological review*, 122–131.

Massey, D. S., & Denton, N. A. (1989). Hypersegregation in us metropolitan areas: Black and hispanic segregation along five dimensions. *Demography*, *26*(3), 373–391.

McPherson, J. M., & Smith-Lovin, L. (1987). Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *American sociological review*, 370–379.

McPherson, M., Smith-Lovin, L., & Brashears, M. E. (2006). Social isolation in america: Changes in core discussion networks over two decades. *American sociological review*, *71*(3), 353–375.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, *27*(1), 415–444.

McRae, K., Cree, G. S., Westmacott, R., & De Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *53*(4), 360.

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

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Murphy, G. L., Chen, S. Y., & Ross, B. H. (2012). Reasoning with uncertain categories. *Thinking & Reasoning*, *18*(1), 81–117.

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive psychology*, *27*(2), 148–193.

Murphy, G. L., & Ross, B. H. (2010a). Category vs. object knowledge in category-based induction. *Journal of Memory and Language*, *63*(1), 1–17.

Murphy, G. L., & Ross, B. H. (2010b). Uncertainty in category-based induction: When do people integrate across categories? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 263–276.

Newell, B. R., Paton, H., Hayes, B. K., & Griffiths, O. (2010). Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin and Review*, *17*(6), 869–874.

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

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

Nosofsky, R. M. (2015). An exemplar-model account of feature inference from uncertain categorizations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1929–1941.

Ostrom, T. M., Carpenter, S. L., Sedikides, C., & Li, F. (1993). Differential processing of in-group and out-group information. *Journal of Personality and Social Psychology*, *64*(1), 21.

Ostrom, T. M., & Sedikides, C. (1992). Out-group homogeneity effects in natural and minimal groups. *Psychological Bulletin*, *112*(3), 536.

Papadopoulos, C., Hayes, B. K., & Newell, B. R. (2011). Noncategorical approaches to feature prediction with uncertain categories. *Memory & cognition*, *39*(2), 304–318.

Park, B., & Hastie, R. (1987). Perception of variability in category development: Instance-versus abstraction-based stereotypes. *Journal of Personality and Social Psychology*, *53*(4), 621–635.

Park, B., & Judd, C. M. (1990). Measures and models of perceived group variability. *Journal of Personality and Social Psychology*, *59*(2), 173.

Park, B., & Rothbart, M. (1982). Perception of out-group homogeneity and levels of social categorization: Memory for the subordinate attributes of in-group and out-group members. *Journal of Personality and Social Psychology*, *42*(6), 1051.

Park, B., Ryan, C. S., & Judd, C. M. (1992). Role of meaningful subgroups in explaining differences in perceived variability for in-groups and out-groups. *Journal of Personality and Social Psychology*, *63*(4), 553.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological bulletin*, *68*(1), 29.

Pickett, C. L., & Brewer, M. B. (2001). Assimilation and differentiation needs as motivational determinants of perceived in-group and out-group homogeneity. *Journal of Experimental Social Psychology*, *37*(4), 341–348.

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. *Current Psychology*, *3*(1), 5–18.

Quattrone, G. A., & Jones, E. E. (1980). The perception of variability within in-groups and out-groups implications for the law of small numbers. *Journal of Personality and Social Psychology*, *38*(1), 141.

Ratcliff, N. J., Hugenberg, K., Shriver, E. R., & Bernstein, M. J. (2011). The allure of status: High-status targets are privileged in face processing and memory. *Personality and Social Psychology Bulletin*, *37*(8), 1003–1015.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition*, *91*(2), 113–153.

Ross, B. H., & Murphy, G. L. (1996). Category-based predictions: Influence of uncertainty and feature associations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 736–753.

Rubin, M., & Badea, C. (2007). Why do people perceive ingroup homogeneity on ingroup traits and outgroup homogeneity on outgroup traits. *Personality and Social Psychology Bulletin*, *33*(1), 31–42.

Rubin, M., & Badea, C. (2010). The central tendency of a social group can affect ratings of its intragroup variability in the absence of social identity concerns. *Journal of Experimental Social Psychology*, *46*(2), 410–415.

Rubin, M., & Badea, C. (2012). Theyre all the same... but for several different reasons a review of the multicausal nature of perceived group variability. *Current Directions in Psychological Science*, *21*(6), 367–372.

Rubin, M., Hewstone, M., & Voci, A. (2001). Stretching the boundaries: Strategic perceptions of intragroup variability. *European Journal of Social Psychology*, *31*(4), 413–429.

Ryan, C. S., Judd, C. M., & Park, B. (1996). Effects of racial stereotypes on judgments of individuals: The moderating role of perceived group variability. *Journal of Experimental Social Psychology*, *32*(1), 71–103.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, *117*(4), 1144.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106.

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, *17*(4), 443–464.

Simon, B., & Pettigrew, T. F. (1990). Social identity and perceived group homogeneity evidence for the ingroup homogeneity effect. *European Journal of*

Social Psychology, *20*(4), 269–286.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, *63*(2), 129.

Stewart, T. L., Vassar, P. M., Sanchez, D. T., & David, S. E. (2000). Attitude toward women's societal roles moderates the effect of gender cues on target individuation. *Journal of Personality and Social Psychology*, *79*(1), 143.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual review of psychology*, *33*(1), 1–39.

Thorndike, E. L. (1927). The law of effect. *The American Journal of Psychology*, *39*(1/4), 212–222.

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(4), 473–479.

Van Kempen, R., & Şule Özüekren, A. (1998). Ethnic segregation in cities: new forms and explanations in a dynamic world. *Urban studies*, *35*(10), 1631–1656.

Verde, M. F., Murphy, G. L., & Ross, B. H. (2005). Influence of multiple categories on the prediction of unknown properties. *Memory and Cognition*, *33*(3), 479–487.

Voci, A., Hewstone, M., Crisp, R. J., & Rubin, M. (2008). Majority, minority, and parity: Effects of gender and group size on perceived group variability. *Social Psychology Quarterly*, *71*(2), 114–142.

Wattenmaker, W. D. (1991). Learning modes, feature correlations, and memory-based categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 908.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological review*, *111*(2), 430.