One and known: Incidental probability judgments from very few samples

Ishan Singhal[1,2] (ishan20@iitk.ac.in), Narayanan Srinivasan[1,2] (nsrini@iitk.ac.in)
[1]Department of Cognitive Science, IIT Kanpur
Kanpur 208016 India
[2]Centre of Behavioural and Cognitive Science, University of Allahabad
Allahabad 211002 India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Department of Computer Science, IIT Kanpur
Kanpur 208016 India

Abstract

We test whether people are able to reason based on incidentally acquired probabilistic and context-specific magnitude information. We manipulated variance of values drawn from two normal distributions as participants perform an unrelated counting task. Our results show that people do learn category-specific information incidentally, and that the pattern of their judgments is broadly consistent with normative Bayesian reasoning at the cohort level, but with large individual-level variability. We find that this variability is explained well by a frugal memory sampling approximation; observer models making this assumption explain approximately 70% of the variation in participants' responses. We also find that behavior while judging easily discriminable categories is consistent with a model observer drawing fewer samples from memory, while behavior while judging less discriminable categories is better fit by models drawing more samples from memory. Thus, our model-based analysis additionally reveals resource-rationality in memory sampling.

Introduction

Statistical inference offers a compelling normative criterion for human judgments during perception and action. Strong evidence of probabilistic reasoning is available for perceptual judgments (Ernst & Banks, 2002) and sensorimotor control (Körding & Wolpert, 2004), wherein human performance in cue integration tasks has been shown to closely follow normative Bayesian principles (Pouget, Beck, Ma, & Latham, 2013). There is also some evidence suggesting how Bayesian updates for such settings could be carried out by populations of neurons (Ma, Beck, Latham, & Pouget, 2006).

Similar normative claims have also been advanced for human cognition, supported by evidence from a a wide variety of cognitive tasks (Oaksford, Chater, et al., 2007; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). However, critics contend that Bayesian analysis offers too many degrees of freedom to the analyst in the design of appropriate priors and likelihood, such that empirical fits with data can often be attributed to environmental and psychological properties embedded in the design of these entities (Jones & Love, 2011).

Properties of the environment and biophysical embodiment can strongly constrain the nature of priors in perceptual and motor control tasks and some language-related cognitive tasks. However, it is harder to produce such constraints for cognitive tasks wherein embodiment is weakly involved and environmental statistics are variable. For instance, to accept Griffiths and Tenenbaum (2006)'s contention that humans are Bayes-optimal in making commonplace probability judgments, we must believe that people *incidentally* keep track of the distribution of a large variety of real-world events. Similarly, to accept the explanation for risk aversion offered in Stewart, Chater, and Brown (2006)'s decision-by-sampling theory, we must believe that people incidentally track the distribution of money magnitudes.

While there is some evidence that people track incidental frequencies well (Hasher & Zacks, 1984), more direct tests of incidental distribution learning have been discouraging (Sailor & Antoine, 2005; Tran, Vul, & Pashler, 2017). Given 180 samples from a bimodal distribution of stimulus locations, for example, humans are effectively random in trying to reproduce the distribution again via sampling (Tran et al., 2017).

In summary, while there is considerable evidence to support the case that humans reason probabilistically given an understanding of the appropriate generative model of the world, it is less clear how they are *incidentally* able to acquire and update these generative models and apply them to novel situations. In this study, we tested peoples' ability to *incidentally* learn about context-specific distributions of magnitudes, and reason with them retrospectively. We analyzed participants' behavior with observer models fitted to observations stochastically identical to the ones participants saw during *incidental* learning to characterize the nature of mental representations used in this retrospective probabilistic reasoning task.

Experiment

We implemented our test for *incidental* learning of context-specific probability distributions as a game where participants were asked to assume the role of financial auditors and check restaurant bills for correct total and tip values (tip values had to be less than 10% of the sum of the items' cost). This simple arithmetic task allowed us to expose participants to the values of the items on the bills, ensuring they paid attention to them while remaining naive to the true purpose of the experiment.

There were two phases in the experiment: *training* and *testing*. During training, participants were shown a number of restaurant bills (see sample bill in Figure 1). Each bill had three items listed with money values. These bills were explicitly identified as being from either "cheap" or "expensive" restaurants. Restaurant category had no bearing on the audit-

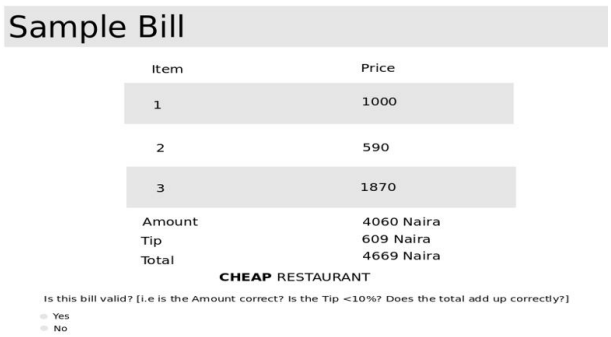

Figure 1: Sample bill presented in the training phase of the experiment.

ing task. The items were denominated in a fictional foreign currency to reduce the influence of participants' prior knowledge of money magnitudes in this context.

The training phase was meant to instantiate the value distributions $p(money|context)$ and the probability of context occurrence $p(context)$ in participants. During the testing phase, on every trial, participants were presented with an item alongside its price, and asked to assess how likely it was that the item came from an expensive restaurant, thus asking them to express $p(context|money)$.

Participants

The study protocol was reviewed and approved by an IRB. The study was conducted using a web interface. A total of 91 (mean age = 25.1years, females = 44) participants completed the experiment. Seven participants with poor performance (less than 80% accuracy) in the training phase were excluded from the study. Unmotivated participants (8 such) who responded with identical responses for all the test phase problems were also excluded from analysis resulting in 76 participants. *A priori* sample size calculated was for 72 participants (Cohen's f = 0.4, power = 0.85). The hypothesis, sample size and analysis plan were formally preregistered (link concealed to preserve anonymity during peer review).

Stimuli

In the training phase of the experiment, we showed participants values drawn from two normal distributions. Values from each distribution were shown as bills from restaurants with two category labels, "cheap" and "expensive" (See Fig.1.). We manipulated the variance (low, medium, high) of these distributions between participants with number of participants being 25, 26 and 25 respectively. Cheap and expensive labels each had their own distribution and we scaled the variance by the mean (fixed means for cheap = 1600 and expensive = 2900, scaling variance factors: low = 0.25, medium = 0.5 and high = 0.7).

Procedure

Training Phase Participants were randomly assigned to one of three variance groups. Thus, a participant could see values drawn in the training phase from either a distribution with low, medium or high variance. The means of these value distributions (for cheap and expensive restaurants) were kept constant across all participants. We asked each participant to audit 20 bills as part of the experiment, with each bill containing three money values sampled from the condition-specific distribution. There were 10 bills each (30 samples; 3 x10) for 'cheap' and 'expensive' restaurant bills. We provided feedback (correct/incorrect) for their responses on each bill as an attention check. Participants who performed poorly on the arithmetic task (accuracy less than 80%) were eliminated from analysis.

Testing Phase During the training phase conducted earlier, participants were unaware that they would be tested on the values given in the bills, i.e. they were tested retrospectively. In this phase, participants were presented with 40 individual items, each with a certain value. They were asked to indicate how likely it is that these items were drawn from an expensive restaurant on a seven point Likert scale. All participants were tested on values generated from the same distribution with variance factor 0.5. This allowed us to compare responses between the three groups and investigate whether their priors for "expensive" and "cheap" restaurants was biased by the variance of the distribution they trained on.

Analysis and Results

We fit each participant's responses from the testing phase with a psychometric function,$\gamma + \frac{\lambda-\gamma}{1+(\frac{x}{c})^{\beta}}$ using the curve fitting toolbox in MATLAB. Here γ was the upper asymptote of the psychometric curve, λ was the lower asymptote, c was the inflexion point of the curve and β was the steepness or the slope of the curve. To fit the curves, we converted Likert ratings to probability judgements by using a recently validated mapping of verbal labels to assigned probabilities (Hancock & Volante, 2020)[1].

We ran a between subject Bayesian ANOVA with three conditions of variance (high, medium, low) for all four parameters in JASP (0.13.1.0) with a default cauchy prior. The alternate hypothesis that variance conditions would change the slopes (β) of the fitted curves had strong evidence (BF_{10} = 7771000, error% = 0.001). Our Post-hoc tests for the same showed that the slopes of the fitted curves for participants in the low variance group were more likely to be higher than the those in the medium (posterior odds = 66391, BF_{10U} = 113025) and high variance group (posterior odds = 1054, BF_{10U} = 1795), while we found no evidence either way for difference in slopes for medium and high variance groups (posterior odds = 1.3, BF_{10U} = 2.28). See figure 2 for the slope differences between the groups. There was weak to moderate evidence for the null for inflexion point (BF = 3.39) and lower bound (BF = 6.26), and inconclusive evidence (BF

[1]Specifically, we mapped our 7-point Likert scale with labels [very unlikely, unlikely, somewhat unlikely, undecided, somewhat likely, likely, very likely] to the numbers [0.05, 0.20, 0.35, 0.50, 0.60, 0.70, 0.90] respectively.

= 0.53) for the upper bound.

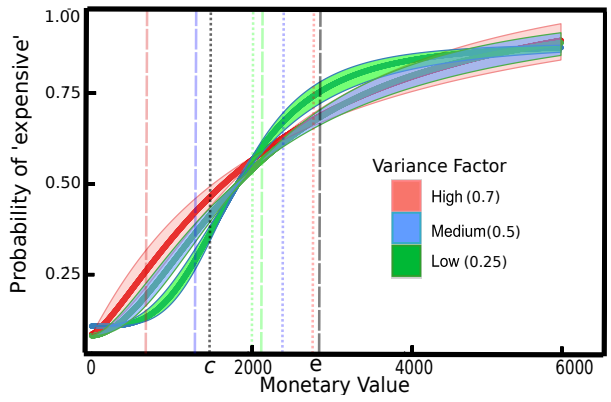

Figure 2: Psychometric function fit using averages of parameters estimated for participants for all three conditions. Shading indicates 95% CI for slope parameter estimates. Dashed and dotted lines indicate mean and SD (one line per condition) for the three 'expensive' and 'cheap' distributions respectively. The color for the shading and dotted lines indicates the three variance conditions.

Discussion

Psychometric analysis of the data supports the hypothesis that the spread of the distribution of numerical observations during training will affect the posterior probability of category judgments elicited during the surprise testing phase. While our experiment is set up to determine whether people can reason with numerical quantities acquired via experience in a Bayesian manner, it can also be viewed as a categorization task. We show observers instances of two categories - cheap and expensive restaurants, and then (stochastically) ask them to classify new instances into one of them. The fact that greater within-category variance makes categorization more challenging is well-documented for both supervised (Alfonso-Reese, Ashby, & Brainard, 2002) and unsupervised category learning (Kloos & Sloutsky, 2008). It is also modeled quite well by classic computational models of categorization (Fried & Holyoak, 1984; Anderson, 1991).

However, *incidental* category learning of the nature we find here has not been documented previously. Participants in our experiment were asked to audit bills for totalling mistakes during training, and were not given feedback about their responses during testing nor did they know that a testing phase would follow. Given the retrospective nature of the design, participants did not know that they had to learn or memorize the values for different category labels. Studies showing successful unsupervised learning of categories tend to use highly separated categories, and any increase in within category variance tends to make category learning very hard even given 100s of training trials (Ell & Ashby, 2012). Viewed in this light, the fact that people are able to *incidentally* learn categories in this task from 30 examples per category is significant[2].Our primary interest in this experiment is to characterize the extent to which participants' behavior can be explained by a Bayesian observer model, which is described in the next section.

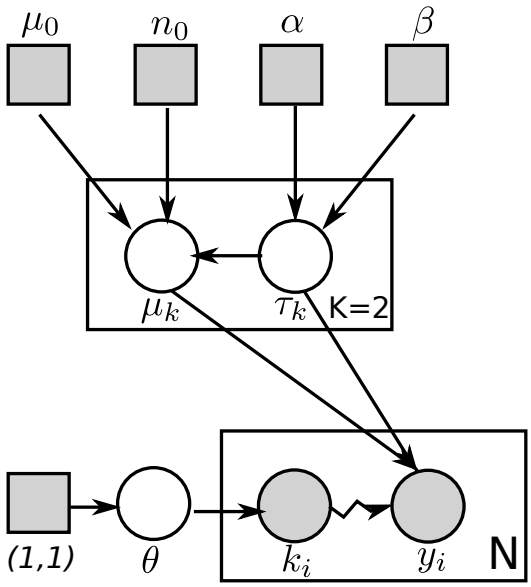

Figure 3: Generative model for experimental task in plate notation.

Ideal observer model

Our observer model is straightforward. From the behavioral results, it is evident that people are able to significantly track the joint distribution $p(K,Y)$, where K is a discrete random variable indexing restaurant category and Y is a continuous random variable indexing natural numbers corresponding to money magnitudes in our experiment. In the testing phase, participants are asked to express $p_{K|Y}(k|y)$, the conditional probability of k, given any particular numerical magnitude y.

The ideal observer will construct this quantity using Bayes rule as follows,

$$p_{K|Y}(k|y) = \frac{p_K(k) f_{Y|K}(y|k)}{f_Y(y)}, \qquad (1)$$

where f_Y is the pdf of a mixture of the two components, and $f_{Y|K}$ is a single component's pdf. To parameterize this model, we model p_K as a Bernoulli trial $p_K(\theta)$, both $p_{Y|K}$ as Gaussian distributions $f_{Y|K}(y|\mu^k,(\tau^k)^{-1})$, and p_Y as a mixture of these Gaussians weighted by the Bernoulli distribution. The distribution $p_{K|Y}$ remains without a concise parametric specification, and is simply readout in tabular form.

We assume that people enter the experiment with some Gaussian $f_{Y|K}$ and f_Y, and update these as they go along in the experiment. We also assume they start with some Bernoulli

[2]See Experiment 4 in Fried and Holyoak (1984) for a historical example of surprising incidental learning of category structure.

p_K and an improper uniform prior of 0.5 on the positive half-line for $p_{K|Y}$. These initial conditions are specified by means of hyper-parameters, following a generative model very similar, but slightly different from a Bayesian Gaussian mixture model (Bishop, 2006), as illustrated in Figure 3. The key difference is that the mixture component identity is co-observed alongside numerical magnitudes on each observation rather than being a latent variable.

When a new observation $\{y_{obs}, k_{obs}\}$ appears, y_{obs} updates the $p_{Y|K}$ distribution corresponding to $K = k_{obs}$, and k_{obs} updates the p_k distribution. We model the k update as a Bernoulli update with a Bayes estimator using an uninformative $Beta(1,1)$ prior.

We model the $y|k$ update as sequential inference about a Gaussian with unknown mean and precision using a normal-inverse-Gamma conjugate prior (Murphy, 2007). The ideal observer begins learning by updating hyperparameters θ for p_K and $\{\alpha, \beta, \mu_0, n_0\}$ for $f_{Y|K}$. Then it updates the marginal f_Y and finally uses Equation 1 to calculate $p_{K|Y}$.

Model-based Analysis

As Jones and Love (2011) point out, it is important to justify the psychological assumptions embedded in Bayesian models to truly bring them into alignment with reality. In the context of our task, the ideal observer assumes that, when asked to express a probability judgment about category membership, a human constructs this belief by retrieving all 60 training samples from memory to update their generative model. Therefore, for our model-based analysis, we also consider two frugal alternatives to this baseline observer model,

1. inspired by prototype models of categorization (Fried & Holyoak, 1984; Minda & Smith, 2001), we consider a *prototype* observer model which assumes that people retrieve the mean of category distributions, and update an abstract generative model of world situations like the given task with just this pair of mental observations while constructing probability judgments in our task.

2. inspired by recent Bayesian sampling proposals (Sanborn & Chater, 2016; Zhu, Sanborn, & Chater, 2020), we consider a memory *sampling* extension of the prototype observer model, which assumes that people either encode or retrieve a subset of training phase observations to construct prototypes, which are then used to construct probability judgments as in the prototype observer model.

We first analyzed all three models' ability to account for our data with a single set of pooled hyperparameters estimated across all participants. We fit the observer models to data by giving it sequential access to 60 observations emitted by the same generative distributions as the original observations seen by each participant in each of the three conditions and then using these three condition-specific trained models to predict posterior probabilities for the 40 observations seen by participants during the test phase. We obtained max likelihood estimates for all parameters, and calculated BIC scores for all models' predictions compared with our data. Since the posteriors emitted by the ideal and sampling observers vary because of the stochastic nature of their likelihoods, we report median BIC scores obtained across 1000 simulations each for these two models.

Model/VF	0.25	0.5	0.7
Ideal	-3085 (11)	-3403 (12)	-3055 (15)
Prototype	-3176	-3434	-3296
Sampling (N=8)	-3194 (6)	-3441 (7)	-3299 (8)

Table 1: BIC scores for observer models fit condition-wise. SEM across multiple runs reported in brackets. All results rounded to integers.

Both the alternative models are clearly superior to the ideal observer across all three training conditions. Interestingly, the sampling observer explains the data better than the ideal observer, as measured by BIC, when implemented with more than 8 samples. Thus, in an empirical sense at least, it is evident that observers that compress incoming magnitude information into category prototype representations are better models for participants' judgments in our task than fully Bayesian observers with perfectly individuated memory of exemplars. The high correspondence between the prototype model and our data (absolute goodness-of-fit $RMSE = 0.15$ across all three conditions) suggests that human observers possess a normatively appropriate generative model applicable for the task we asked them to do retrospectively and invert it to obtain situation-specific probabilistic judgments by drawing a few samples of past observations from memory.

We next considered model fits with hyperparameters estimated at the per participant level. Figure 4A-B summarizes goodness-of-fit for the ideal observer making predictions on the observations seen by each participant during the test phase. Panel A plots the model's predicted posterior distribution $p_{K|Y}$ against one participant's probabilistic responses. Panel B summarizes median goodness-of-fit, quantified using adjusted R^2 obtained for participants, training the model using the same training distributions as each participant in a given condition, and setting hyperparameters for each participant using max likelihood estimation. The results suggest that the ideal observer models variability in individuals' responses for easier discrimination judgments reasonably, but is increasingly less effective in modeling variability in responses for harder discriminations.

Figure 4C-D summarizes goodness-of-fit for these two observer models. Overall, they afford better descriptions of individuals' data than the ideal observer model, because of the extra modeling flexibility afforded them by the introduction of two (prototype model - both prototype values) and one (sampling model - sample count) additional parameters.

An additional analysis of the sampling observer model affords additional insight into peoples' behavior. Figure 5 plots adjusted R^2 values across all participants achieved by sampling models restricted to drawing a fixed number of obser-

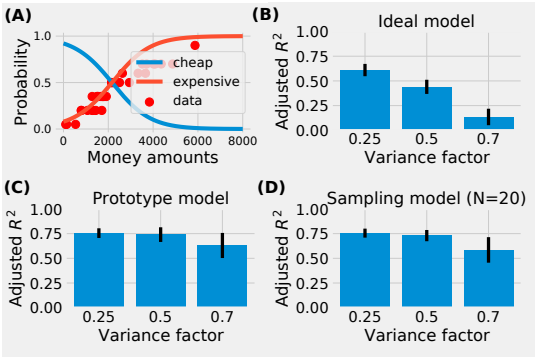

Figure 4: Ideal observer model (A) fit to one participant's data, and (B) median goodness-of-fit measured across all participants by condition. Median goodness of fit for a (C) prototype observer model and a (D) sampling observer model. Error bars represents ± 1 s.e.m.

vations from memory per category. The value reported in the graph is the median of 100 runs of the model for each sample count level to account for the stochasticity of memory sampling. We see that the model's ability to explain behavioral variation for participants exposed to high variance training distributions continues to improve with the number of additional samples permitted. However, interestingly, drawing as few as one to two samples per category is already sufficient for the sampling observer model to explain more than 60% of the variance in peoples' behavior if they were assigned to the low variance training condition.

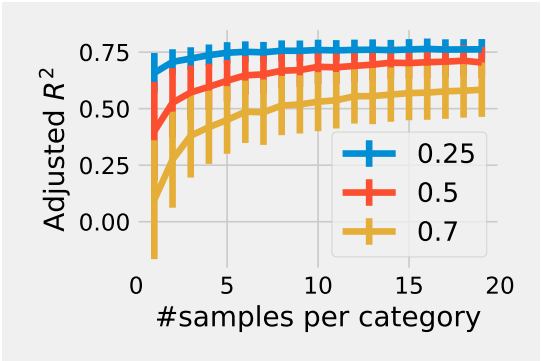

Figure 5: Median sampling observer model R^2 measured across 100 iterations each for different counts of observations sampled. Iteration R^2 value is calculated as the average R^2 of models with that sample count fit to each participants' data. Error bars show average 1 s.e.m. across iterations.

The general trend seen in Figure 5 suggests that people might be constructing prototypes using fewer samples from memory for categories that are less confusable, and more samples for categories that are more confusable. This pattern is consistent with people performing resource-rational memory sampling - drawing fewer samples from memory to construct prototypes needed to discriminate easily discriminable categories, and drawing more samples to construct prototypes needed to discriminate less discriminable categories. Resource-rational meta-reasoning has several attractive theoretical properties (Griffiths, Lieder, & Goodman, 2015), but opinion remains divided on the extent to which such resource-rationality is simply a metaphor to add greater flexibility to Bayesian models (Rahnev, 2020), or whether it can be shown to have stronger ontic commitments (Griffiths et al., 2015). The pattern of meta-reasoning in Figure 5 supports the consideration of sampling as a concrete operationalization of resource-rationality for Bayesian models of cognition (Srivastava & Schrater, 2014; Sanborn & Chater, 2016).

General Discussion

We make three contributions in this paper. One, we present empirical evidence of incidental supervised category learning from very few observations, consistent with a long-standing assumption in Bayesian cognitive science that people intuitively even when not explicitly motivated store important statistical information about events in the world (Griffiths & Tenenbaum, 2006; Stewart et al., 2006). Two, we demonstrate using computational modelling, that humans' behavior in tests of incidental category learning are consistent with the behavior of a Bayesian observer that samples a subset of event observations from memory to construct probability judgments (Srivastava & Schrater, 2014; Zhu et al., 2020). Three, we find that behavior while judging easily discriminable categories is consistent with a model observer drawing fewer samples from memory, while behavior while judging less discriminable categories is better fit by models drawing more samples. These observations are consistent with theoretical expectations of resource-rationality in memory sampling (Griffiths et al., 2015; Sanborn & Chater, 2016).

Our work also demonstrates quite clearly that, whether people are able to incidentally learn distributions corresponding to real-world situations (Hasher & Zacks, 1984) or not (Tran et al., 2017), they can certainly learn enough about them well enough from as few as one sample per distribution to reason probabilistically about them, as presupposed by earlier work (Griffiths & Tenenbaum, 2006; Stewart et al., 2006). Interestingly, the prototype observer model shows excellent empirical fits with data across our three variance manipulations, even though the model itself gains no access to variance information. In short, incidental probability judgments are consistent with a 'one-and-known' sampling heuristic supporting peoples' probabilistic causal inference.

Probabilistic accounts of cognition are sometimes questioned on grounds of neurobiological plausibility. Probabilistic population codes can potentially encode and enable reasoning over distributions corresponding to invariant or slow-changing priors and well-structured likelihoods (Ma et al., 2006). It is possible to model important elements of perceptual and motor tasks using such distributions, but not for cognitive tasks of the nature we consider in this paper. Flex-

ible probabilistic reasoning in everyday cognition appears to require operations incongruent with the extent to which and timescales on which synaptic weights can change as a function of experience (Malinow, Madison, & Tsien, 1988). Our results offer a possible resolution for this problem - sampling a small number of observations from memory, is sufficient to tune a generic situation model, potentially learned over long experience, into expressing probabilistic judgments that closely match human behavior.

Due to the key role played by the observer's possession of the correct generative model in explaining behavior with very little further learning, this work also amplifies an important open question. How can we characterize the set of situations for which people possess such generative models, and situations for which they don't? While proposals for learning generative models by induction exist (Kemp & Tenenbaum, 2008; Tenenbaum et al., 2011), considerable work remains to characterize the library of generative models that people carry around in their heads.

Acknowledgements

This work was supported by a CSRI, DST India grant DST/CSRI/2017/334

References

Alfonso-Reese, L. A., Ashby, F. G., & Brainard, D. H. (2002). What makes a categorization task difficult? *Perception & Psychophysics*, *64*(4), 570–583.

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

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Ell, S. W., & Ashby, F. G. (2012). The impact of category separation on unsupervised categorization. *Attention, Perception, & Psychophysics*, *74*(2), 466–475.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429–433.

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(2), 234.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in cognitive science*, *7*(2), 217–229.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological science*, *17*(9), 767–773.

Hancock, P., & Volante, W. G. (2020). Quantifying the qualities of language. *PloS one*, *15*(5), e0232198.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American psychologist*, *39*(12), 1372.

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and brain sciences*, *34*(4), 169.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kloos, H., & Sloutsky, V. M. (2008). What's behind different kinds of kinds: effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature neuroscience*, *9*(11), 1432–1438.

Malinow, R., Madison, D. V., & Tsien, R. W. (1988). Persistent protein kinase activity underlying long-term potentiation. *Nature*, *335*(6193), 820–824.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Murphy, K. P. (2007). Conjugate bayesian analysis of the gaussian distribution. *def*, *1*($2\sigma 2$), 16.

Oaksford, M., Chater, N., et al. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170–1178.

Rahnev, D. (2020). Resource-rational analysis vs. resource-rational humans. *The Behavioral and brain sciences*, *43*, e19.

Sailor, K. M., & Antoine, M. (2005). Is memory for stimulus magnitude bayesian? *Memory & Cognition*, *33*(5), 840–851.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in cognitive sciences*, *20*(12), 883–893.

Srivastava, N., & Schrater, P. (2014). Frugal preference formation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1–26.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *science*, *331*(6022), 1279–1285.

Tran, R., Vul, E., & Pashler, H. (2017). How effective is incidental learning of the shape of probability distributions? *Royal Society Open Science*, *4*(8), 170270.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*.