Is it Time Bayes went Fishing?
Bayesian Probabilistic Reasoning in a Category Learning Task

Marcus Lindskog (marcus.lindskog@psyk.uu.se), Anders Winman (anders.winman@psyk.uu.se), Peter Juslin (peter.juslin@psyk.uu.se)
Department of Psychology, Uppsala University, P. O Box 1225,
Uppsala, 751 42 Sweden

Abstract

People have generally been considered poor at probabilistic reasoning, producing subjective probability estimates that far from accord to normative rules. Features of the typical probabilistic reasoning task, however, make strong conclusions difficult. The present study, therefore, combines research on probabilistic reasoning with research on category learning where participants learn base rates and likelihoods in a category-learning task. Later they produce estimates of posterior probability based on the learnt probabilities. The results show that our participants can produce subjective probability estimates that are well calibrated against the normative Bayesian probability and are sensitive to base rates. Further, they have accurate knowledge of both base rate and means of the categories encountered during learning. This indicates that under some conditions people might be better at probabilistic reasoning than what could be expected from previous research.

Keywords: Probabilistic reasoning, category learning, Bayes' theorem, base rate

Introduction

Research concerned with human probability judgment has been dominated by the general conclusion that people are poor at reasoning with probabilities because they substitute hard facts about probabilities with subjective variables that are conveniently available (see e.g., Gilovich, Griffin, & Kahneman, 2002). In fact, with respect to tasks requiring people to integrate probabilities according to Bayes' theorem the verdict is even harder, as summarized by a quote from Kahneman and Tversky (1972, p. 450): "*In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not a Bayesian at all.*" In the present study, we present results indicating that, at least under some conditions, the claim by Kahneman and Tversky might have been somewhat premature.

To appreciate the kind of task our participants are faced with, imagine going to catch fish in a lake where the fishing authorities have farmed two kinds of bass: copper bass, and silver bass. The two kinds of bass look identical and the only feature that distinguishes them is that the copper bass weighs, on average, a little less than the silver bass. While looking identical they, however, taste very differently. If you want a delicious dinner, you should go for the silver bass while if you want to feel sick you should choose a copper bass. To make sure that the lake is not over fished the authorities have also decided that, at all time, the ratio of copper to silver bass should be 8:2, a piece of information not made publically available.

The fish scenario illustrates a type of situation that people engage in frequently in their everyday lives. The fishermen estimate the probability of a new object belonging to a category based on previous experience. That is, each time a fish is taken out of the lake the fisherman needs to estimate the probability of a given fish being a copper or a silver bass. The estimate is informed by experience with fish previously taken up out of the lake and cooked for dinner, thus effectively categorized as copper or silver bass. More specifically, this illustrates a situation where an observer needs to learn *base rates* and *likelihoods* from experience and later integrate this information to reach an estimate of a posterior probability. In such, the fish scenario incorporates two areas of cognitive psychology: *probabilistic reasoning* and *category learning*, that have been extensively investigated separately, but seldom together (but see, Nilsson, Olsson, & Juslin, 2005).

Probabilistic Reasoning

Research on human probabilistic reasoning has mainly been concerned with the evaluation of subjective probability estimates against normative rules of probability. In the typical experiment, the subjective estimates are informed by a set of probabilities explicitly stated in the task. Consider, for example, the *cab problem* (Tversky & Kahneman, 1980) where participants are asked to estimate the probability of a cab involved in an accident being blue rather than green based on the base rates of blue (.15) and green (.85) cabs and the hit-rate (.8) of an eyewitness with both the base rate and hit-rate being explicitly stated in the task. The normative answer (.41) can be found by integrating the information in the problem using Bayes' theorem.

When presented with the cab problem, and similar problems, people tend to give probability estimates that are much higher than what is implied by Bayes' theorem. Often the modal response is closer to the hit-rate of the eyewitness (.8). This pattern of results is commonly interpreted as a captivation in participants by the hit-rate along with neglect of the base rate (.15). The dominating explanation to this apparent neglect of base rates has been that people are prone to use judgmental heuristics (e.g. the representativeness heuristic) that ignore base rates (e.g., Kahneman & Tversky, 1972; but see, Koehler, 1996). More recent accounts of probabilistic reasoning, suggesting that people are prone to linear additive information integration, argue instead that the non-normative answers are the result of how probabilities are integrated rather than the use of heuristics per se (Juslin, Nilsson, & Winman, 2009; Juslin, Nilsson, Winman, & Lindskog, 2011).

Regardless of the underlying mechanisms explaining the results, the use of complex normative rules, such as Bayes' theorem, to integrate probabilities seems to be beyond the ability of most people (e.g., Eddy, 1982; Gigerenzer & Hoffrage, 1995). In fact, even explicit instructions regarding how to use Bayes' theorem to integrate the information is insufficient to improve people's judgments (Juslin et al., 2011). It should be noted, however, that the despite the somewhat discouraging picture painted by previous research, recent accounts of human cognition (e.g., Oaksford & Chater, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) have indicated that people are rational Bayesian agents with a remarkable ability to integrate information in accordance with the laws of Bayesian probability theory.

The extent to which people's probability estimates in Bayesian reasoning tasks coincide with the normative answer has largely been tested using tasks similar to the cab problem. Three features of these types of tasks are noteworthy, features that might influence the conclusions that can be drawn about human probabilistic reasoning. First, the information to be integrated (base rates, likelihoods, etc.) is explicitly given to participants in the form of probabilities (e.g., Kahneman & Tversky, 1972) or, sometimes, frequencies (e.g., Gigerenzer & Hoffrage, 1995). Second, the tasks are commonly set up to give a posterior probability that is low, often .40 or smaller. Finally, the outcome for which the posterior probability is estimated is often binary (blue or green cab, disease or no disease, engineer or lawyer, etc.). All of these task features make it difficult to draw strong conclusions about the ability of people to integrate probabilistic information. In everyday life, people are unlikely to come across situations where probabilities are explicitly stated. They rather encounter situations, like the fishing example above, where probabilities are learned from experience. Many real life situations also include an outcome, for which the posterior probability is estimated, that is continuous rather than binary. Furthermore, the restriction of the range of posterior probabilities makes conclusions about the extent to which people are calibrated against the Bayesian probability difficult due to regression effects. In order to address these three issues it is necessary to find a task where participants learn probabilities from experience and where it is possible to elicit probability estimates on the entire 0 to 1 range for a continuous outcome variable. One promising candidate is found in category learning.

Probabilistic Reasoning and Category learning

In the typical categorization task participants are presented with a number of stimuli from two or more categories and are asked to assign an appropriate category to each based on a set of features. During learning, the categorization is often followed by feedback regarding the correct category.

The literature contains several different models of how categorization is made, including prototype, exemplar, and decision-bound models (Ashby & Maddox, 2005). The purpose of this study is not to distinguish between the different kinds of models. Rather, we draw upon the notion that most models of human categorization make assumptions about: a) how and what information is accessed from the categories and what computations are performed on this information and, b) how a response is selected after computations are made (Ashby & Alfonso-Reese, 1995). For most models that assume a probabilistic, in contrast to a deterministic, response selection process, the decision rule subjects are assumed to use could be described as; respond category A to stimulus *x* with probability *M*(x) where:

$$M(x) = \frac{\beta_A S_{xA}}{\beta_A S_{xA} + \beta_B S_{xB}}. \qquad (1)$$

In this expression β_i is the response bias towards category *i* and S_{xi} is a measure of the similarity between stimulus *x* and category *j*. At least under some conditions Eq. 1 can be reduced to

$$M(x) = \frac{\hat{P}(A)\,\hat{f}_A(x)}{\hat{P}(A)\,\hat{f}_A(x) + \hat{P}(B)\,\hat{f}_B(x)}, \qquad (2)$$

where $\hat{P}(i)$ and $\hat{f}_i$ are estimators of the base rate and probability density function of category *i* respectively (Anderson, 1991; Ashby & Alfonso-Reese, 1995). Ashby and Alfonso-Reese (1995) argued that these properties of the categorization task transform it into a density estimation task where participants are faced with estimating base rates and probability density functions of each category. Indeed, several investigations of models of categorization have shown that they are mathematically equivalent to density estimation (e.g., Anderson, 1991; Ashby & Alfonso-Reese, 1995; Griffiths, Sanborn, Canini, & Navarro, 2008)

The similarities between Bayes' theorem and Eq. 2 suggest that categorization tasks are similar to probabilistic reasoning tasks with the difference that while probabilities are explicitly stated in the reasoning task they need to be learned from trial-by-trial feedback in the categorization task. Further, while the literature on probabilistic reasoning is somewhat pessimistic about people's ability to integrate probabilities the categorization literature suggests that people are quite apt at categorization (Ashby & Maddox, 2005). However, while research on categorization has been extensively concerned with how categories are represented and the processes leading up to a categorization (Ashby & Maddox, 2005) it has put much less focus on the extent to which base rates and likelihoods are learned. Further, the typical categorization task requires participants to assign a stimulus to a category leaving the question of whether M(x) in Eq. 2 is close to the normative posterior probability unanswered.

It should be noted that categorization research indicates that people are able to learn base rate information from experience (Medin & Edelson, 1988), at least under some conditions, and that models of categorization can be seen as the cognitive substrate of subjective probability estimates (Nilsson et al., 2005).

The Present Study

The present study investigates the accuracy of subjective probability estimates in a Bayesian probability reasoning task. Instead of being presented with base rates and likelihoods explicitly, however, participants learn them through experience in a categorization task.

Further, we elicit probabilities from the entire range of possible posterior probabilities for a continuous outcome variable in order to have a task that is as ecologically valid as possible.

To investigate factors that might influence the learning of base rates and likelihoods as well as the process used to elicit probability estimates, we manipulate both base rate and the distance between categories (i.e., the likelihood ratio).

Method

Participants

Participants were 40 (24 female and 16 male) undergraduate students from Uppsala University with a mean age of 25.1 years (*SD* = 4.3 years). They received a movie ticket or course credits for their participation.

Design

The experiment used a 2x2 between-subjects design with base-rate-ratios (8:2/6:4) and category-distance (short / far) as independent variables.

Materials and Procedure

The computerized task was carried out on a PC and consisted of a learning phase and a test phase. On each of the 200 trials in the learning phase, participants categorized an exemplar to one of two categories (A and B) along a single dimension. The number of exemplars from each category was determined by the base-rate-ratio. In the 8:2-condition the ratio of the number of exemplars in the two categories was 8:2 (i.e., 160 A-exemplars and 40 B-exemplars) and in the other condition it was 6:4. The 200 items were presented in an individually randomized order.

A unique training set was created for each participant by randomly sampling stimuli from two Gaussian distributions with equal standard deviation ($\sigma = 6$). In the short category-distance condition, the mean of the two Gaussians were 40 and 49 respectively while in the far condition they were 40 and 52. Whether category A or B had the highest mean was counterbalanced over participants.

The experiment used two cover stories. Either the categories where two types of projectors (Braun / Kodak) categorized on their brightness (lumens) or two types of disease (Buragamo / Terrigitis) categorized on the fictitious PKS-value. Participants were told that the values they would experiences were created specifically for this study and that they could not use any prior knowledge to solve the categorization task. The two cover stories, and which category was A or B, was counterbalanced over participants.

On each of the 52 trials in the test phase participants were presented with a value (lumens or PKS) not seen in training and were asked to state the probability (in percent) that the item belonged to category A (i.e., the category with the highest base rate). To create the 52 items for the test phase the range of the training set was divided into eleven intervals based on the posterior probability p_{Ax} that a test item x belonged to category A ($p_{Ax} = 0$, $0 < p_{Ax} < .1$, $.1 \leq p_{Ax} < .2$, …, $.9 \leq p_{Ax} < 1.0$, $p_{Ax} = 1.0$). For each of the nine middle intervals ($0 < p_{Ax} < .1$, … $9 \leq p_{Ax} < 1.0$) four items were randomly drawn uniformly from that interval. Six items each were randomly drawn from the two extreme intervals, where the posterior probability is 0 and 1. Finally, four critical items with an equal distance to the category means were included in the test set. After completing the test phase participants gave explicit estimates of the base rates and means of the two categories.

Results

Learning Performance

To investigate learning performance, the learning phase was divided into 10 blocks of 20 trials each. For each block, we calculated the proportion of correct categorizations. Figure 1 illustrates that participants quickly learn to categorize the training stimuli to the appropriate category with proportion correct reaching .8 at the end of the training phase.

We investigated the extent to which the base-rate-ratio and category-distance manipulations influenced the rate of learning by entering proportion correct as dependent variable into a 2x2x10 split plot ANOVA with base-rate-ratio (8:2 / 6:4) and category-distance (short / far) as between-subjects independent variable and training block as within-subjects independent variable. The analysis revealed a significant main effect of training block ($F(9, 324) = 4.95$, $MSE = 0.012$, $p < .001$) with a significant difference between the first and last block. Further, there was a significant main effect of category-distance ($F(1, 36) =$

5.09, *MSE* = 0.068, *p* = .03) where participants in the far condition performed better (*M* = .78, *SEM* = .018) than participants in the short condition (*M* = .72, *SEM* = .018). Notably this difference was significant also in the last training block (*t*(38) = 2.6, *p* = .01).

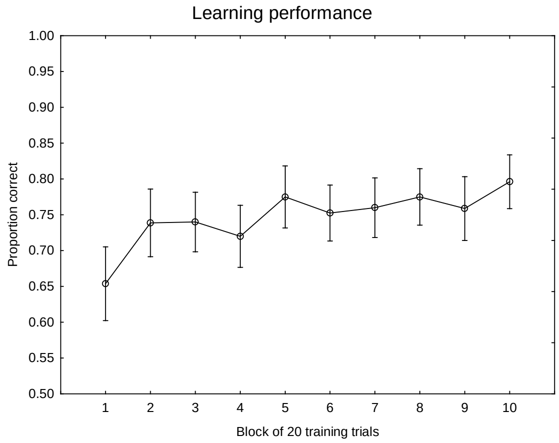

Figure 1: Proportion correct as a function of training block. Vertical bars denote 95 % - confidence intervals.

Neither the main effect of base-rate-ratio (*F*(1, 36) = 2.83, *MSE* = 0.068, *p* = .10) nor any of the interactions (all *p*:s > .13) reached significance. Notice that while the main effect of category-condition indicates that it was easier for participants to learn the categories with means far apart as opposed to close together, the lack of interactions suggest an equal learning rate in all conditions.

Subjective Probability Estimates

In the test phase participant gave explicit estimates of the posterior probability that an item *x* belongs to category A (i.e., the category with the highest base rate). Figure 2 shows the mean estimated probability plotted against the normative Bayesian probability. In the figure, estimates are grouped into the eleven intervals described above.

As is evident from the figure participants are on average fairly well calibrated in their subjective probability estimates. To investigate the effect of base-rate-ratio and category-distance on the subjective estimates of posterior probability we calculated the mean absolute difference between the estimated and normative probability. The difference was entered as dependent variable into a 2x2 factorial ANOVA with base-rate-ratio (8:2 / 6:4) and category-distance (short / far) as between-subjects independent variables. There were no significant effects (all *p*:s > .18). Thus, probability estimates were on average not influenced by base-rate-ratio or category-distance.

To investigate a possible bias in the probability estimates the signed difference (rather than absolute difference) was entered into the corresponding ANOVA. Once again there were no significant effects (all *p*:s > .26) and a single sample t-test on the signed difference revealed that it did not differ significantly from 0 (*t*(39) = .96, *p* = .35).

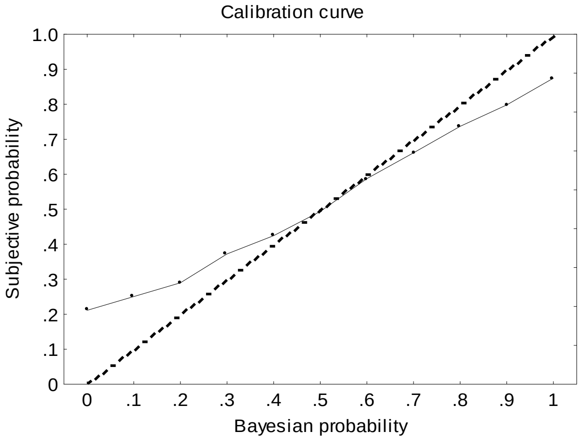

Figure 2: Subjective probability plotted against the normative Bayesian probability. Dotted line indicates perfect calibration.

The results illustrated in Figure 2 indicate that the accuracy of subjective probability estimates might vary as a function of the Bayesian posterior probability. To investigate this probability we conducted a more fine grained analysis where Bayesian probability interval was added as a within-subjects factor in the analysis of absolute error. This 2x2x11 split-plot ANOVA revealed two significant effects. First, the main effect of Bayesian probability interval was significant (*F*(10, 360) = 3.07, *MSE* = 0.018, *p* < .001). The effect is due to absolute errors for the larger probability intervals being smaller than those for the lower intervals. Second, the significant probability interval by base-rate-ratio (*F*(10, 360) = 2.79, *MSE* = 0.018, *p* < .001) is illustrated in Figure 3 by means of a calibration curve. As can be seen in the figure, the interaction is due to estimates in the low probability intervals being slightly better for participants in the 6:4-condition than for participants in the 8:4-condition while it is the opposite in the high probability intervals.

The analysis above suggests that the base-rate-ratio manipulation might influence the extent to which participants use base rates to inform their subjective probability estimates. To investigate this possibility we analyzed participants’ probability estimates of the critical items included in the test set. Remember that the critical items are positioned with the same distance to both category means. If participants disregard the base rate information and instead use the ratio of the distance from a test item to each of the two means as a proxy for the posterior probability, or some similar strategy, they should estimate the posterior probability of all critical items to be .5.

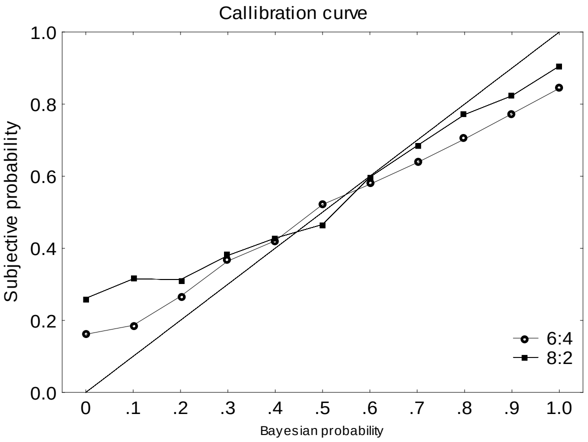

Figure 3: Subjective probability plotted against the normative Bayesian probability for the two base-rate-ratio conditions separately. Dotted line indicates perfect calibration.

Figure 4 displays the distribution of responses to the critical items. As is evident from the figure a majority of responses are larger than .5, indicating that participants take the base rate of the two categories into account when giving subjective probability estimates. To further investigate the use of base rates the subjective probability estimates of critical items were entered as dependent variable into a 2x2 factorial ANOVA with base-rate-ratio and category distance as between-subjects factors. One participant, considered an outlier (|z| > 2.5), was excluded from the analysis. The ANOVA revealed a significant main effect of base-rate-ratio (*F*(1, 35) = 4.63, *MSE* = 0.037, *p* = .038) with higher probability estimates in the 8:2-condition (*M* = .76, *SD* = .14) then in the 6:4-condition (*M* = .62, *SD* = .24). None of the other effects reached significance (both *p*:s > .20). More importantly in all conditions, participants gave estimates larger than .5, even though not significantly larger in the short-6:4-condition, indicating sensitivity to base rates.

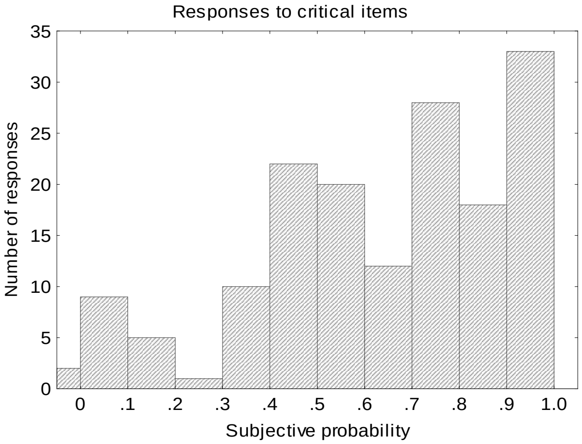

Figure 4: Distribution of subjective probability estimates of critical items in the test phase.

A further indication of sensitivity to base rates is given by the explicit estimates of base rates illustrated in Figure 5.

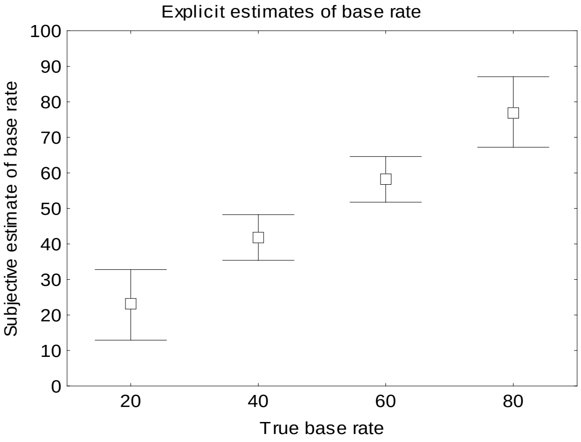

Figure 5: Means of explicit estimates of base rate for the four different true base rates separately. Vertical bars denote 95 % - confidence intervals.

As can be seen in the figure the explicit estimates are sensitive to the experienced base rates. In addition there is little difference in the accuracy of estimates in the different conditions indicating that the differences in use of base rates seen above is not an effect of differences in learning.

Discussion

Research on probabilistic reasoning has long been dominated by the general conclusion that people are very poor at integrating information according to the laws of probability (e.g., Bayes' theorem). At the same time research concerned with category learning, indicates that people are quite apt at solving categorization tasks that, at least mathematically, are similar to probabilistic reasoning tasks. In the present study, we therefore combined these two research traditions by eliciting subjective posterior probabilities from base rates and likelihoods learned in a categorization tasks.

Performance in the learning phase indicated that our participants quickly learned to categorize the stimuli correctly. Performance was somewhat better when category means were far apart as opposed to close together. This was expected because the closer the two category means get the more two their probability density functions overlap, which in turn makes it more difficult to distinguish the two categories.

The subjective probability estimates given by participants in the test phase were, as is illustrated in Figure 2, well calibrated against the normative Bayesian probability. There was no systematic bias in the estimates and the pattern of results seen in Figure 2 suggests that the deviations from the normative Bayesian probability could be attributed to regression effects. Notably, even though there was a difference in learning between the two category-distance conditions, this did not affect the correspondence of the subjective estimates.

The explicit estimates of base rates and category means indicated that participants learned these category properties. Arguably, however, they might not have used them to reach a subjective probability estimate. However, the analysis of the critical items included in the test phase showed that participants in all conditions were sensitive to the base rate and, at least to some extent, integrated this knowledge in their probability estimates.

Similar to previous research demonstrating that people can be good at reasoning under some conditions (e.g., Baron, 2000), the results of the present study show that when people are allowed to learn base rates and likelihoods in a category learning task they are at least under some conditions able to produce subjective probability estimates that are well calibrated and sensitive to base rates. This suggests that the conclusion by Kahneman and Tversky (1972, p. 450) may have been somewhat premature. An interesting question for future research is to investigate the processes leading up to what is apparently a normative answer.

Acknowledgments

This research was sponsored by the Swedish Research Council.

References

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

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

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56*, 149–178.

Baron, J. (2000). Can we use human judgments to determine the discount rate? *Risk Analysis*, *20*, 861-868.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Herustics and Biases* (pp. 249–267). Cambridge, UK: Cambridge University Press.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684–704.

Gilovich, T., Griffin, D. W., & Kahneman, D. (Eds.). (2002). *Heuristics and Biases: The psychology of intuitive judgment*. New York: Krieger Publishing Company.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* (pp. 303–328). Oxford, UK: Oxford University Press.

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*, 856–874.

Juslin, P., Nilsson, H., Winman, A., & Lindskog, M. (2011). Reducing cognitive biases in probabilistic reasoning by the use of logarithm formats. *Cognition*, *120*(2), 248–267.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1–53.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68.

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 600-620.

Oaksford, M., & Chater, N. (2009). Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences*, *32*, 69–120.

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