Where Do Probability Judgments Come From? Evidence for Similarity–Graded Probability

Peter Juslin (peter.juslin@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Håkan Nilsson (hakan.nilsson@97.polmag.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Henrik Olsson (henrik.olsson@psy.umu.se)
Department of Psychology, Umeå University
SE-901 87, Umeå, Sweden

Abstract

This paper compares four models of the processes and representations in probability judgment. The models represent three principles that have been proposed in the literature: 1) the *representativeness heuristic* (interpreted as relative likelihood or prototype-similarity), 2) *cue-based relative frequency*, and 3) *similarity-graded probability*. An experiment examined if these models account for the probability judgments in a category learning task. The results indicated superior overall fit for similarity-graded probability throughout training. In the final block, all models except similarity-graded probability were refuted by data.

Introduction

Where do probability judgments come from? This question has been fiercely debated the last decades in research on judgment under uncertainty. In the late sixties the conclusion was that probability judgments are fairly accurate reflections of *extensional* properties of the environment such as frequencies (Peterson & Beach, 1967). This changed with the influential *heuristics and biases* program in the seventies and eighties, which emphasized that probability judgments are guided by *intensional* aspects like similarity (Kahneman, Slovic, & Tversky, 1982). The nineties saw a renewed interest in the idea that extensional properties are reflected in peoples' probability judgments as specified by the *ecological models* (Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994). A third alternative combines intensional and extensional properties in an exemplar model to produce *similarity-graded probabilities* (Juslin & Persson, 2000).

Only rarely have these accounts been contrasted in studies that chart the processes and representations that underlie probability judgments. We compare four models of how people make probability judgments in a category learning task. The task involves assessment of the probability that a probe with feature pattern t belongs to one of two mutually exclusive categories, A or B. For example, a physician may assess the probability that a patient with symptom pattern t suffers from one of two diseases. The models represent three principles that have been proposed in the judgment literature: the *representativeness heuristic* (two versions), *cue-based relative frequency*, and *similarity-graded probability*. We present a category structure that allows us to contrast predictions derived from these hypotheses.

Representativeness Heuristic

According to the representativeness heuristic, people judge the probability that an object or event belongs to a category on the basis of the degree to which it is representative of the category, or reflects salient features of the process that generated it (Kahneman et al., 1982). The representativeness heuristic is routinely evoked post hoc to explain cognitive biases but has not been subjected to careful tests in inductive learning tasks.

A *relative-likelihood* interpretation of representativeness states that the probability judgment $p(A)$ that probe t belongs to A is made by comparing the likelihood of t in category A relative to its likelihood in categories A and B:

$$p(A) = \frac{.5d + f(t|A)}{d + f(t|A) + f(t|B)}, \qquad \text{Eq. 1}$$

where $f(t|A)$ and $f(t|B)$ are the *relative frequencies* of feature patterns identical to t in categories A and B, respectively. To allow for pre-asymptotic learning (Nosofsky, Kruschke, & McKinley, 1992) and response error in the use of the overt probability scale (Erev, Wallsten, & Budescu, 1994), all models in this paper are equipped with a free parameter d for dampening. The dampening effectively pulls the predictions towards .5 (e.g., an un-dampened prediction of 1 becomes somewhat less extreme as a result of d). Eq. 1 implies that the probability judgment that, say, a patient with symptom pattern t has disease A is a direct function of

the likelihood of these symptoms given disease *A*.[1]

A *prototype* interpretation of representativeness is that the probability judgments derive from the *similarities* $S(t|P_A)$ and $S(t|P_B)$ of *t* to the category prototypes P_A and P_B, respectively:

$$p(A) = \frac{.5d + S(t|P_A)}{d + S(t|P_A) + S(t|P_B)}, \qquad \text{Eq. 2}$$

where the similarity is computed by the multiplicative similarity rule of the context model (Medin & Schaffer, 1978),

$$S(t, y) = \prod_{j=1}^{D} d_j \ , \ d_j = \begin{cases} 1 \ if \ t_{.j} = y_{.j} \\ s \ if \ t_{.j} \neq y_{.j} \end{cases}, \qquad \text{Eq. 3}$$

where *y* is a prototype (as in Eq. 2 above) or an exemplar (as in Eq. 5 below). The value of d_j is 1 if the values on feature *j* match and *s* if they mismatch. *Similarity s* is a free parameter in the interval [0, 1] for the impact of mismatching features.

On this view, the probability judgment that a patient with symptom pattern *t* has disease *A* is a function of *t*'s similarity to the prototypical symptom pattern for disease *A*. The prototype is defined by the modal (i.e., most frequent) feature value in the category on each feature dimension. When the feature values are equally common, we selected the feature value that generated the more frequent overall pattern in the category.

Cue-Based Relative Frequency

The idea that probability judgments derive from cue-based relative frequency is represented by *Probabilistic Mental Model theory (PMM-theory;* Gigerenzer et al., 1991; see e.g., Juslin, 1994, for similar ideas). These ideas have been used to scaffold global predictions in studies of realism of confidence, but not been tested in studies of inductive learning.

In the current context, we interpret PMM-theory as suggesting that the probability judgment that probe *t* belongs to category *A* is a function of the cue value (α_1) of the single most valid cue that can be applied:

$$p(A) = \frac{.5d + F(A|\alpha_1 ***)}{d + F(A|\alpha_1 ***) + F(B|\alpha_1 ***)}, \qquad \text{Eq. 4}$$

where $F(A|\alpha_1***)$ and $F(B|\alpha_1***)$ are the *frequencies* of category *A* and *B* exemplars with cue value α_1, respectively, and the symbol "∗" denotes that the other cue values are discarded (there are four features in the experiment presented below). Eq. 3 represents the relative frequency of category *A* conditional on presence of cue value α_1. Thus, a subjective probability judgment is a reflection of the validity of the cue with the highest cue-validity that is present in the event or object being judged. This strategy is known as *Take The Best* (TTB) meaning that you rely on the cue with the highest validity (Gigerenzer, Todd, & the ABC Group, 1999).

Similarity-Graded Probability

A class of models that combines intensional and extensional aspects is exemplar models in categorization research. In exemplar models, decisions are made by comparing new objects with exemplars stored in memory. The *context model* (Medin & Schaffer, 1978) responds to both similarity (intensional property) and frequency (extensional property) in general, and to only one of these factors in predictable circumstances (Juslin & Persson, 2000). PROBEX (i.e., PROBabilities from EXemplars; Juslin & Persson, 2000) is a model of probability judgment based on the context model.

With PROBEX, probability judgments are made by comparisons between the probe *t* and retrieved exemplars x_i (i = 1…I). The exemplars are represented as vectors of *D* features (in the present experiment, *D*=4 and the features are binary). Continuing with the example of medical diagnosis, a patient with symptom pattern *t* leads to retrieval of stored exemplars of previous patients with similar symptoms and their diagnoses. The probability judgment is a weighted average of the outcome indices $c(x_i)$ for the exemplars, where $c(x_i)$=1 for exemplars in category *A* and $c(x_i)$=0 for exemplars in category *B*. The weights in the average are the respective probe-exemplar similarities $S(t|x_i)$:

$$p(A) = \frac{.5d + \sum_I S(t|x_i)c(x_i)}{d + \sum_I S(t|x_i)}, \qquad \text{Eq. 5}$$

where similarity is computed from Eq. 3. This hypothesis implies that if a new patient with symptom pattern *t* is similar to many exemplars x_i with diagnosis *A*, the probability that the new patient has disease *A* is high.

The complete version of PROBEX involves a sequential sampling of exemplars, but this aspect is ignored in the present application. This effectively reduces Eq. 5 to the original context model (Medin & Schaffer, 1978) with a dampening (see Nosofsky et al., 1992, for a similar formulation), but with one crucial difference: $p(A)$ does not refer to a predicted proportion of category *A* classifications, but to a prediction of a probability judgment.

With similarity parameter *s*=0, only exemplars with feature patterns identical to *t* affect the judgment and Eq. 5 emulates a "*picky frequentist*" (Juslin & Persson, 2000).[2] Ignoring the dampening *d*, Eq. 5 then computes

[1] "Direct function" means that the predicted probability judgments are a function of likelihoods alone, not likelihoods *and* prior probabilities, as implied by Bayes' theorem.

[2] This version of Eq. 5 is formally identical to Bayesian estimation of a probability with the Beta-distribution and parameters α and β equal to $.5d$.

the relative frequency of category *A* among exemplars with identical features. For $s>0$, Eq. 5 computes a *similarity-graded probability* that is both affected by the frequency of exemplars, and the probe-exemplar similarities. Note that, although PROBEX responds to similarity, it is not identical to the representativeness heuristic. For example, PROBEX (Eq. 5) cannot produce a conjunction fallacy, unless amended with auxiliary assumptions of some sort (Juslin & Persson, 2000). PROBEX has been fitted to people's probability judgments in a general knowledge task (Juslin & Persson, 2000) but not been tested in inductive learning tasks.

Category Structure and Predictions

The problem with contrasting these three hypotheses is that in most category structures, they generate highly correlated predictions. Table 1, however, provides one category structure that implies qualitatively distinct predictions for certain critical exemplars (Figure 1).

Table 1: The categories with the 20 x 3 exemplars.

X	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	Category		
1	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
2	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
3	1	1	1	1	3	3	3	3	5	5	5	5	A	A	A
4	1	1	1	1	3	3	3	3	5	5	5	5	A	A	B
5	1	1	1	1	3	3	3	3	5	5	5	5	A	B	B
6	1	0	0	0	3	2	2	2	5	5	5	4	A	A	A
7	1	0	0	0	3	2	2	2	4	4	4	4	A	A	B
8	1	0	0	0	3	2	2	2	4	5	4	4	A	A	B
9	0	0	0	0	3	2	2	2	4	4	4	4	A	A	B
10	0	0	0	0	3	2	2	2	4	5	4	4	A	A	B
11	1	1	0	0	3	2	2	2	4	4	4	4	B	A	B
12	1	1	0	0	3	2	2	2	4	4	5	4	B	A	B
13	0	1	0	0	3	2	2	2	4	4	4	4	B	A	B
14	0	1	0	0	3	2	2	2	4	4	5	4	B	A	B
15	0	1	0	0	3	2	2	2	4	4	4	4	B	A	B
16	0	0	1	1	3	2	2	2	4	4	4	5	B	A	B
17	0	0	1	1	3	2	2	2	4	4	4	4	B	A	B
18	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B
19	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B
20	0	0	1	1	2	3	3	3	4	4	4	4	B	B	B

The design involves 60 exemplars with four features each, organized into three substructures. The 20 exemplars in the first substructure have features C_1-C_4, the 20 in the second substructure have features C_5-C_8 and the last 20 have features C_9-C_{12}. The feature has two possible values (0 vs. 1, for C_1-C_4; 2 vs. 3 for C_5-C_8; 4 vs. 5 for C_9-C_{12}). The last three columns headed by "Category" specify whether the exemplar is in category *A* or *B*. The first column is for exemplars with features C_1-C_4, the second for exemplars with features C_5-C_8, and the third for exemplars with features C_9-C_{12}.

In the first part of the experiment, the 60 exemplars are presented with feedback about whether they belong to category *A* or *B*. In the second part, the participants are asked to estimate the probability that probes with certain feature patterns belong to category *A*. There are *fifteen distinctive feature patterns*, six for features C_1-C_4, three for features C_5-C_8, and six for features C_9-C_{12}. The participants estimate the probability of category *A* for all fifteen patterns. The critical patterns are 1111 for features C_1-C_4, 3333 for C_5-C_8 and 5555 for C_9-C_{12}. Across these, the models provide distinctly different predictions (see Figure 1).

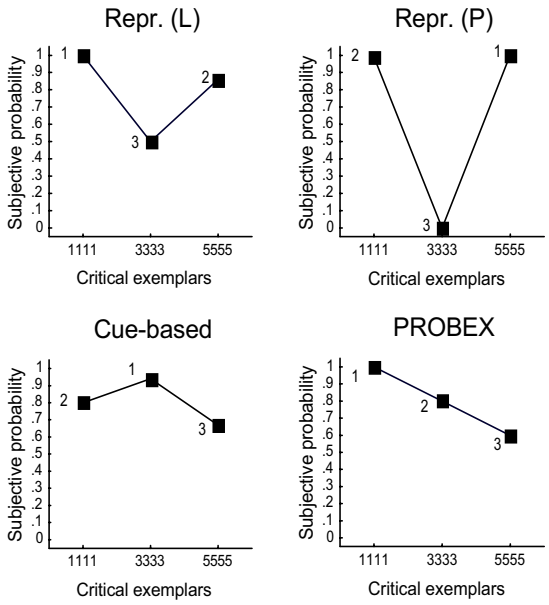

Figure 1: Predicted probability judgments. All predictions are derived with d=0. The predictions for representativeness with prototype similarity (P) are based on s=.1. The predictions for PROBEX are based s=0 (i.e., Picky frequentist).

For example, the predictions for feature pattern 3333 are derived as follows. The representativeness heuristic with a likelihood interpretation implies $p(A)$ = .25/(.25+.25) = .5: the probe is identical to 25% of the exemplars in category *A* and 25% of the exemplars in category *B*. In regard to a representativeness heuristic with prototype similarity, we note that the prototypes for category *A* and *B* in the second substructure (i.e., based on C_5-C_8) are 3222 and 2333, respectively. Ignoring the dampening *d*, Equation 2 implies the prediction $s^3/(s^3+s)$. The prototype for *A* differs on three features and the prototype for *B* on one feature. The prediction depends on the parameter *s*, but it will generally be low and always lower than .5. With cue-based relative frequency, $p(A)$ = 16/(16+1) = .94. Given the value of 3 for the most valid cue C_5, 16 of 17 exemplars belong to category *A*. According to the picky frequentist prediction by PROBEX (s=0), $p(A)$ = 4/5 = .8. Four out of five exemplars with identical feature patterns belong to

category *A*. At *s*>0, the prediction falls below .8. Predictions for the other two critical patterns are derived in the same way.

Note in Figure 1 that, depending on the model, the probability judgments for the three critical patterns have a different *rank order.* These predicted rank orders are *a priori* and not dependent on the parameters (i.e., *s* or *d*). By comparing the observed with the predicted rank order, we get a qualitative test of the models. In addition, we can evaluate the quantitative fit of the models to the judgments for all 15 feature patterns.

Method

Participants

Twenty-four undergraduate students (10 men and 14 women) in the age of 19 to 32 (average age = 23.3) participated. The participants where paid between 65-86 SEK depending on their performance. They received 30 SEK plus 1 SEK for each correct answer in the last learning block.

Apparatus and Materials

The experiment was carried out on a PC–compatible computer. In each of the four training blocks, the program first presented the 60 exemplars from Table 1. The task involved judgments for 60 companies, where 20 companies belonged to each of three countries (substructures). Each exemplar had four features that differed depending on the country. The features are presented in Table 2. The features and names of the countries were chosen to be as neutral as possible. In the test phase after each training block, the program presented each of the 15 distinct feature patterns twice.

Design and Procedure

A two-way within-subjects design was used. The independent variables were the number of training blocks (four blocks) and category substructure (three substructures). The dependent variable was the probability judgments. The specific assignment of concrete cue labels (see Table 2) to the abstract category structure (see Table 1) was varied and counterbalanced across the participants. Thus, each concrete label in Table 2 appeared equally often in each of the three substructures and equally often in the role of each of the abstract features denoted C_1 to C_{12} in Table 1.

The participants were to act as stockbrokers assigned to invest a large sum of money in three countries about which they knew nothing. They were told that it is usually enough to know four company features to know if the stock will rise or fall in the next twelve-months, but that the features differ between the countries.

Table 2: Twelve concrete features used in the experiment.

Features	Descriptions
1) Listed at the LAP / IPEK stock exchange?	
2) Less / more than 1000 employees?	
3) Commercials on television / the radio?	
4) Changed owner / merged in last three years?	
5) Less than / more than three years old?	
6) Give money to charity / sponsor sports team?	
7) Active in specific region / whole country?	
8) Co-operation with university / own research department?	
9) In state-financed SKATOS / TAPOS program?	
10) Primarily export-based / import-based?	
11) Affirmative action based on gender / ethnic background?	
12) Stock risen / fallen during the last 12-month?	

The participants were told that the first phase is a training session where they are presented with 60 companies, each described by four features that depend on the country. The features describe the companies as they were twelve months ago. They were to guess whether the stocks rose (*A*) or fell (*B*) in value in the last year. After each judgment, they received feedback on the actual development. The four features were presented on the screen. Below the question "Will the stock-value rise or fall during the next twelve month?" appeared. The participant answered s (short for the Swedish word for rise) or f (short for the Swedish word for fall). Thereafter, the correct answer appeared together with the company's four features.

In the test phase, the participants were told that they were to see a set of companies as they are today and judge the probability of an increase in their stock-value and that the markets are identical on all parameters today as they were one year ago. The feature patterns were presented in the same way as in the training phase, but with the question: "What is the probability that the stock of this company increases in value in the next 12 months?" They were told to answer in percentages and even up to 0, 10…100.

The test blocks consisted of two assessments of the 15 distinct feature patterns, one for rising stock-value (*A*) and one for falling stock-value (*B*). This allowed us to examine the additivity of the probability judgments (i.e., if the mean probability assigned to *A* and *B* for a feature pattern sum to 1). To get reliable data we recoded probability-*B* judgments into probability-*A* judgments by subtracting the probability-*B* judgments from 1. There was no feedback. The order of the probability judgments was counterbalanced within participants. The training and test blocks were repeated four times. The entire procedure took between one hour and fifteen minutes to two hours.

Results

Figure 2 presents mean probability judgments for the critical feature patterns in each of the four test blocks. The data for the third block shows a tendency to agree with the prediction by PROBEX. The fourth block exhibits clear agreement with the prediction by PROBEX. The confidence intervals for exemplars, 1111 and 5555 are clearly separated and the predicted decreasing trend is observed which refute all models except PROBEX.

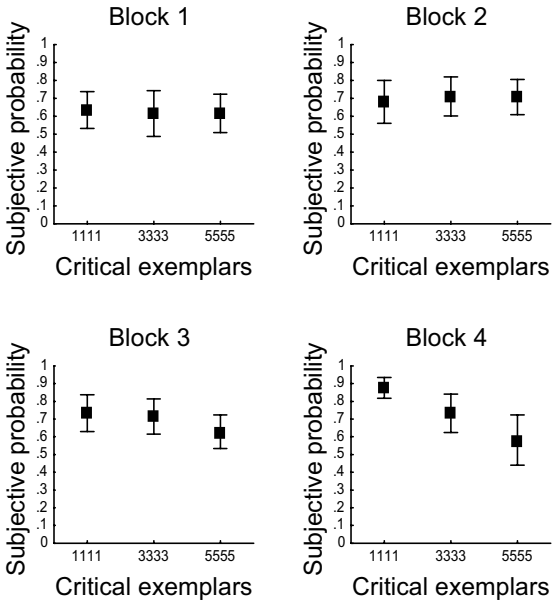

Figure 2: Means with 95% confidence intervals for the estimations of the critical exemplars for the four test blocks.

For the first two blocks, the data reveal no clear trend favoring any of the four models. One tentative interpretation of this result is that it reflects a mix of individual strategies in the early stages of training. To explore this more carefully, we fitted the four models to the data from all 15 distinct feature patterns. The probability judgments proved to be additive on average (i.e., the mean probability assigned to *A* and *B* for a feature pattern sum to 1).

The models were fitted to the mean probability judgments for each of the 15 distinct feature patterns with Root Mean Square Deviation (RMSD) as error function. This was done separately for each of the four test blocks. The model based on the representativeness heuristic as relative likelihood has one free parameter (*d*), representativeness heuristic as prototype similarity has two free parameters (*s* & *d*), cue-based relative frequency has one free parameter (*d*), and exemplar-based retrieval (PROBEX) has two free parameters (*s* & *d*). The results are summarized in Table 3.

Table 3 verifies that in the later stages of training, PROBEX provides a good fit to the data. Because the standard error of measurement is .05, the RMSDs for PROBEX (.054 & .058) come close to saturating the data. Considering all four blocks it is clear that cue-based relative frequency fits the judgments poorly in all blocks. Although the qualitative pattern in Figure 1 for blocks 1 and 2 does not accord with PROBEX, we find that it is the best fitting model throughout training. The models based on the representativeness heuristic exhibit moderate fit early in training, which successively deteriorates with training.

Table 3: Fit of the models as a function of test block in terms of RMSD and coefficients of determination r^2.

		Test Block			
Model	Index	1	2	3	4
Repr. (L)	RMSD	.087	.111	.105	.124
	r^2	.65	.69	.70	.73
Repr. (P)	RMSD	.094	.123	.124	.158
	r^2	.61	.62	.58	.55
Cue-based	RMSD	.139	.193	.188	.234
	r^2	.20	.21	.23	.22
PROBEX	RMSD	.060	.067	.054	.058
	r^2	.87	.92	.92	.95

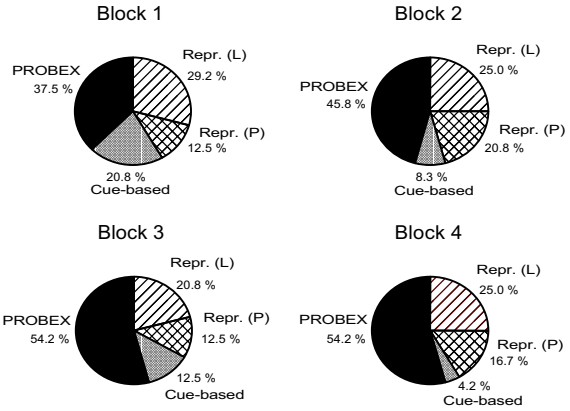

Figure 3: The percent of participants best described by each of the four models, in each of the four test blocks.

Finally, these conclusions were verified at the level of individual participants. The same model-fitting procedure was performed for each participant, with the exception that all models were fitted with one free parameter (*d*). In each block, the percentage of participants for which each model provided the best fit was ascertained. Figure 3 shows that PROBEX is the most frequent winner, although a minority of participants is

better fitted by representativeness as relative likelihood, mostly in the early test blocks.

Discussion

Research on subjective probability judgment has been characterized by a normative stance, where judgments are compared to norms from probability theory. Cognitive theory has primarily been evoked to provide post hoc explanations, as in most applications of the representativeness heuristic, or as scaffolds for more general predictions, as in the applications of cue-based relative frequency. The point of departure for our research is the need to make closer contact between cognitive theory and judgment research in controlled studies that allow us to support or refute core concepts in judgment research, such as the representativeness heuristic.

The results reported here provide clear support for the hypothesis of similarity-graded probability (Juslin & Persson, 2000). That an exemplar model is successful may not appear surprising considering the impressive performance of exemplar models in categorization studies (Nosofsky & Johansen, 2000). Yet, the results are at variance with crucial ideas in judgment research, like that of a representativeness heuristic (Kahneman et al., 1982) or cue-based relative frequency (Gigerenzer et al., 1991; Juslin, 1994).

The second to best fitting model was representativeness as relative likelihood, but this may be spurious as, the crucial feature patterns in Figure 1 aside, the predictions by the models tend to be correlated. However, the superiority of PROBEX is not a mere consequence of a greater inherent flexibility. To demonstrate this, we used the predictions for the last test block by representativeness as relative likelihood as fictive "true data" and added a normally distributed random error with a standard deviation of .05 to mimic measurement error. To this fictive data set, representativeness provided a superior fit (RMSD=.053, r^2=.97) as compared to PROBEX (RMSD=.096, r^2= .83). Thus, the better fit of PROBEX appears to reflect more than larger flexibility in the face of random error.

The best-fitting version of PROBEX (s=.21) in the last test block is not the Picky frequentist version identical to Bayesian estimation of the probability with a Beta-distribution (see Footnote 2). This suggests that, at least in regard to this more simplistic implementation of a Bayesian algorithm, PROBEX provides a better fit to data.

The main objection against the present study is perhaps that it is a single study involving one specific category structure. The category structure used here was guided by the aim of allowing qualitatively distinct predictions by the four models. This category structure may accidentally favor one model over another. Perhaps, a category structure more coherently organized around prototypes yields more support for representativeness as prototype similarity? Likewise, a more feature-rich category structure that posits more demand on information search may yield more support for cue-based relative frequency in the form of TTB (Gigerenzer et al., 1999). Only further research can tell. In any event, these hypotheses will have to count with a serious contestant in the form of PROBEX.

Acknowledgments

Bank of Sweden Tercentenary Foundation supported this research.

References

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101,* 519-527.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probalistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*, 506-528.

Gigerenzer, G., Todd, P. M, & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Juslin, P. (1994). The overconfidence phenomenon as a consequence of informal experimenter-guided selection of almanac items. *Organizational Behavior and Human Decision Processes, 57,* 226-246.

Juslin, P., & Persson, M. (2000). *Probabilities from exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge*. Manuscript submitted for publication.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning, *Psychological Review, 85*, 207-238.

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7,* 375-402.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211-233.

Peterson, C. R, & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68*, 29-46.