ELSEVIER

Cognitive Science 26 (2002) 563–607

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

PROBabilities from EXemplars (PROBEX): a "lazy" algorithm for probabilistic inference from generic knowledge

Peter Juslin [a,*], Magnus Persson [b]

[a] *Department of Psychology, Umeå University, SE-901 87 Umeå, Sweden*
[b] *Department of Psychology, Uppsala University, Box 1225, SE-751 42 Uppsala, Sweden*

Received 14 February 2002; received in revised form 30 May 2002; accepted 30 May 2002

Abstract

PROBEX (PROBabilities from EXemplars), a model of probabilistic inference and probability judgment based on generic knowledge is presented. Its properties are that: (a) it provides an exemplar model satisfying bounded rationality; (b) it is a "lazy" algorithm that presumes no pre-computed abstractions; (c) it implements a hybrid-representation, *similarity-graded probability*. We investigate the *ecological rationality* of PROBEX and find that it compares favorably with Take-The-Best and multiple regression (Gigerenzer, Todd, & the ABC Research Group, 1999). PROBEX is fitted to the point estimates, decisions, and probability assessments by human participants. The best fit is obtained for a version that weights frequency heavily and retrieves only two exemplars. It is proposed that PROBEX implements speed and frugality in a psychologically plausible way.

Keywords: PROBEX; Lazy algorithm; Probabilistic inference

1. Introduction

A common way to address probability judgment in *Cognitive Science* is by asking people to make confidence or subjective probability judgments in regard to general knowledge beliefs. For example, to assess the probability that one city has a larger population than another (e.g.,

* Corresponding author. Tel.: +46-90-786-64-25; fax: +46-90-786-66-95.
E-mail address: peter.juslin@psy.umu.se (P. Juslin).

0364-0213/02/$ – see front matter
PII: S0364-0213(02)00083-6

Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994), or that a briefly described person belongs to an occupational category (Kahneman, Slovic, & Tversky, 1982). These tasks require on-the-spot elaboration of generic knowledge originally acquired for other purposes, in other contexts, by means of *probabilistic inference* to produce a *subjective probability*. Yet, few models address the issue of how subjective probabilities (or "degrees of belief" in the words of Ramsey, 1931) are computed from generic knowledge.

In the 1960s, research emphasized that probability judgments are fairly accurate projections of *extensional* properties of the environment (e.g., frequencies) and thus accord approximately with normative models (Peterson & Beach, 1967). In the 1970s, it was stressed that probability judgments were (mis)guided by *intensional* aspects, like similarity, and therefore prone to cognitive bias (Kahneman et al., 1982).

In this article, we propose that *exemplar models* (e.g., Medin & Schaffer, 1978; Nosofsky, 1984; Kruschke, 1992) provide a fertile approach to probabilistic inference, and that they afford a principled understanding of the role of frequency and similarity in human judgment. A main objection to exemplar models is that they make unreasonable demands on storage and retrieval capacity (Barsalou, Huttenlocher, & Lamberts, 1998; Nosofsky, Palmeri, & McKinley, 1994). As emphasized by Gigerenzer et al. (1999), cognitive algorithms need to make psychologically plausible demands on time, knowledge and computation: they need to implement *bounded rationality* (Simon, 1990).

With this background in mind, we investigate the cognitive processes that compute subjective probabilities from generic knowledge, a topic addressed in terms of three interrelated issues. *First*, we present the original *context model* of perceptual classification (Medin & Schaffer, 1978) and illustrate why it is a useful tool for investigating the role of similarity and frequency in judgment. The context model specifies conditions that emphasize similarity or frequency, respectively, and implements a hybrid-representation, *similarity-graded probability*, that strikes a useful compromise between relevance and sufficient sample size. This allows for efficient exploitation of limited knowledge of an environment.

Second, we present a modification of the context model, *PROBEX* (PROBabilities from EXemplars) that applies to inference from generic knowledge. In addition to the refinements required for this new application (e.g., to generate probability judgments), the modifications aim to produce an exemplar model that satisfies bounded rationality. We investigate the ecological rationality of PROBEX by comparing it to linear multiple regression and two fast-and-frugal algorithms (Gigerenzer et al., 1999). We emphasize that, in contrast to its competitors, PROBEX relies on no pre-computed abstractions and thus belongs to a class of *lazy algorithms* introduced in the artificial intelligence literature (Aha, 1997).

The final section examines the fit of PROBEX to the point estimates, decisions, probability judgments, and response times by human participants performing a paradigmatic example of a generic knowledge task. We demonstrate that the best-fitting version of PROBEX implements a fast-and-frugal exemplar model that makes inferences by the speeded retrieval of a few highly similar exemplars. We conclude that a fast, frugal, and lazy exemplar algorithm like PROBEX provides a more plausible account of the probabilistic inferences that people make in generic knowledge tasks than previous accounts in terms of cue-based inference (e.g., Gigerenzer et al., 1991; Juslin, 1994).

2. Similarity-graded probability

In the last 15 years, one trend in *Cognitive Science* has been a move from abstractions (i.e., rules, schemes, prototypes) to concrete experiences. Theories that stress the storage of concrete instances (traces, exemplars) have appeared in research on memory (Hintzman, 1984, 1988), perceptual categorization (Estes, 1994; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1984, 1986; Nosofsky & Palmeri, 1997; Kruschke, 1992), expertise (Reisbeck & Schank, 1989), automatisation (Logan, 1988), judgment (Kahneman & Miller, 1986), decision making (Klein, 1989), social cognition (Smith & Zarate, 1992), and function learning (DeLosh, Busemeyer, & McDaniel, 1997), and other areas.

The perhaps most successful of these models is the *context model* of perceptual classification (Medin & Schaffer, 1978), later developed into the *generalized context model* (GCM) for continuous dimensions (Nosofsky, 1984, 1986). GCM has been amended with sequential sampling mechanisms in the decision process (Nosofsky & Palmeri, 1997) and the build-up of the stimulus representation (Lamberts, 2000). *ALCOVE* (Kruschke, 1992), which conjoined GCM with the architecture of a neural network, was combined with *generalized recognition theory* (Ashby & Townsend, 1986) to arrive at a model embodying both rules and exemplars (Erickson & Kruschke, 1998). It seems fair to conclude that the context model has been successful, both in accounting for data and in stimulating novel theorizing.

The development presented in this article—PROBEX—is framed in terms of the original context model for binary features (Medin & Schaffer, 1978).[1] In order to apply it to probabilistic inference from generic knowledge, it is amended in the following respects. (a) Features of a presented probe are retrieved from memory rather than extracted from a visual input. (b) Retrieval of exemplars is sequential and terminated by a stopping rule. This avoids exhaustive retrieval of exemplars, but also allows prediction of response times (Nosofsky & Palmeri, 1997). (c) Response rules for point estimates and subjective probability assessments are added. (See Andersson & Fincham, 1996; DeLosh et al., 1997; Dougherty, Gettys, & Ogden, 1999; Smith & Zarate, 1992, for similar proposals in other applications.) We defer a detailed discussion of the relationship between PROBEX and other models to Section 6.

Before we present PROBEX, we highlight the properties that make the context model particularly useful for addressing the issue of whether subjective probabilities derive from representations of environmental frequencies (e.g., Gigerenzer et al., 1991; Juslin, 1994) or similarity relations (Tversky & Kahneman, 1983).

2.1. The original context model

Consider a participant who has stored N_A exemplars $\bar{a}_i$ ($i = 1, \ldots, N_A$) from Category A and N_B exemplars $\bar{b}_i$ ($i = 1, \ldots, N_B$) from Category B. (The bars over a and b denote that they are vectors.) Each exemplar is represented by a vector of D binary feature values $\bar{x}_i = [x_{i1}, x_{i2}, \ldots, x_{iD}]$, where feature value 1 denotes presence of the feature and 0 denotes absence of the feature. Table 1 illustrates six exemplars with four features (i.e., $N_A + N_B = 6$; $D = 4$) where, for example, exemplar 1 [1, 1, 1, 1] has all four features. The participant is presented with a novel probe $\bar{t}$ and is required to categorize it. The similarity $S(\bar{t}, \bar{a}_i)$ between probe $\bar{t}$ and exemplar $\bar{a}_i$ and the similarity $S(\bar{t}, \bar{b}_i)$ between probe $\bar{t}$ and exemplar $\bar{b}_i$ is computed

Table 1
Six exemplars with four features each, distributed into two categories, Larry and Barry

Exemplar	Feature				Category	Similarity to Mr. X
	1	2	3	4		
1	1	1	1	1	*Larry*	1
2	1	1	1	1	*Larry*	1
3	1	1	1	1	*Barry*	1
4	1	1	1	0	*Barry*	s
5	1	1	0	0	*Barry*	s^2
6	1	0	0	0	*Barry*	s^3

The predicted probability assessment P(Larry) that a new probe Mr. X [1, 1, 1, 1] is named *Larry* is computed for three values of the similarity parameter s.

1. Picky probability that Mr. X [1, 1, 1, 1] belongs to Category *Larry* ($s = 0$):

$$P(\text{Larry}) = \frac{2 \times 1^4}{2 \times 1^4 + (1^4 + 0 + 0 + 0)} = \frac{2}{3}.$$

2. Sloppy probability that Mr. X [1, 1, 1, 1] belongs to Category *Larry* ($s = 1$):

$$P(\text{Larry}) = \frac{2 \times 1^4}{2 \times 1^4 + 4 \times 1^4} = \frac{1}{3}.$$

3. Similarity-graded probability that Mr. X [1, 1, 1, 1] belongs to Category *Larry* ($s = .5$):

$$P(\text{Larry}) = \frac{2 \times 1^4}{2 \times 1^4 + (1^4 + .5 + .5^2 + .5^3)} = \frac{2}{3.875} = .52.$$

by the multiplicative similarity rule of the context model,

$$S(\bar{t}, \bar{y}) = \prod_{j=1}^{D} d_j, \quad d_j = \begin{cases} 1, & \text{if } t_j = y_j, \\ s, & \text{if } t_j \neq y_j, \end{cases} \tag{1}$$

where $\bar{y}$ could be any exemplar in Category A or B, d_j is 1 if the values on feature j match and s if they mismatch, s is a parameter in the interval [0, 1] for the impact of mismatching features. Although the context model allows for separate similarity parameters for each feature dimension, in this article we will settle with a single value of s.

From Eq. (1) we get a similarity to each exemplar. This similarity determines its activation in memory and thus its impact on the classification. The original context model implies that the probability P(A) of a classification of probe $\bar{t}$ in Category A is computed by summation of the activation across Categories A and B.

$$P(\text{A}) = \frac{\sum_{i=1}^{N_\text{A}} S(\bar{t}, \bar{a}_i)}{\sum_{i=1}^{N_\text{A}} S(\bar{t}, \bar{a}_i) + \sum_{i=1}^{N_\text{B}} S(\bar{t}, \bar{b}_i)}. \tag{2}$$

One interesting property of exemplar models is that they respond to both frequency and similarity. GCM thus account for the effects of similarity and frequency on the *classification probabilities* in perceptual categorization tasks (Nosofsky, 1988). For the moment, we suppress

the distinction between classification probability and probability judgment and emphasize two special cases of Eq. (2). Later, we make this distinction explicit and present separate response-rules for classification decisions and probability judgments.

2.1.1. Environment parameters

First, if the similarities between probe $\bar{t}$ and all exemplars in Categories A and B is constant, Eq. (2) simplifies to a pure response to frequency (i.e., similarity is a constant that can be factored out to cancel in the nominator and denominator):

$$P(\mathrm{A}) = \frac{N_\mathrm{A}}{N_\mathrm{A} + N_\mathrm{B}}. \tag{3}$$

Eq. (3) corresponds to the relative frequency estimates modeled by the *combined error model* (Juslin, Olsson, & Björkman, 1997; Juslin, Wennerholm, & Olsson, 1999). Eq. (3) represents the extensional feature of relative frequency and the output tends naturally to conform to the probability calculus. Eq. (3) holds in an environment with homogenous objects as, for example, when a participant observes the turn-over of cards from a deck only containing red and blue cards. Eq. (3) might suggest that "normative models provide a good first approximation" to behavior (Peterson & Beach, 1967, p. 42).

If, on the other hand, the number N of exemplars in Categories A and B, respectively, are the same ($N = N_\mathrm{A} = N_\mathrm{B}$), Eq. (2) simplifies to a pure response to similarity (i.e., because $\sum_{i=1}^{N} S(\bar{t}, \bar{x}_i) = N m_{S_x}$, N can be factored out to cancel in nominator and denominator)

$$P(\mathrm{A}) = \frac{m_{S_\mathrm{A}}}{m_{S_\mathrm{A}} + m_{S_\mathrm{B}}}, \tag{4}$$

where m_{S_A} and m_{S_B} are the mean similarities between the probe $\bar{t}$ and the exemplars in Categories A and B, respectively. In an environment with objects that vary in similarity, Eq. (4) has some resemblance to the *representativeness heuristic* (e.g., Kahneman et al., 1982; Tversky & Kahneman, 1983), the alleged cause of a number of cognitive biases. For example, the judgments from Eq. (4) will not respond to the reliability of the information, and thus not be sufficiently regressive (Kahneman et al., 1982). However, in contrast to the representativeness heuristic, Eq. (4) does not produce the conjunction fallacy or base-rate neglect unless conjoined with auxiliary assumptions. We return to these issues in Section 6 when we discuss MINERVA-DM (Dougherty et al., 1999).

The algorithm is the same in both Eqs. (3) and (4)—which are merely special cases of Eq. (2)—and the parameter s may be the same in both cases. Because of the differences in the learning environment, observable behavior is nevertheless predicted to "change" into a response solely to frequency or solely to similarity. In most environments, behavior is affected by both frequency and similarity as governed by the similarity parameter s.

2.1.2. Organism parameter

Table 1 provides the example of a person with six friends, all of which happen to be named either *Larry* or *Barry*. Each friend is described by four features. With the context model, this is modeled by six exemplars, two from the category Larry and four from the category Barry, where exemplars called Larry are homogenous with regard to the four features. Obviously,

the similarity and the frequencies of the exemplars in Table 1 differ, so behavior is potentially sensitive to both similarity and frequency. The person next encounters Mr. $\bar{t}$ modeled by the vector [1, 1, 1, 1]. The person is asked to assess the subjective probability P(Larry) that the name of Mr. $\bar{t}$ is Larry, assuming that Larry and Barry is the only possible choice. How is this assessment affected by the value of the parameter s?

First, consider the possibility that s is 0. We then have the case of a "*picky frequentist*," who only activates exemplars that are identical to the probe. Therefore, the assessed probability P_{picky}(Larry) will equal 2/3, the relative frequency of *Larry*-exemplars within the reference class of exemplars identical to Mr. $\bar{t}$ (computational example 1 in Table 1). Next, consider if s equals 1. In this case we end up with a "*sloppy frequentist*," to whom all exemplars seem equally similar to Mr. $\bar{t}$ (computational example 2). The sloppy frequentist will thus respond with the overall base-rate of Larry-exemplars, 1/3. For all values of s between 0 and 1, the person will respond with a *similarity-graded probability*, where the exemplars are weighted by their similarity (computational example 3).

A hypothesis motivating the research with PROBEX is that similarity-graded probability solves an adaptive problem. Consider the classical problem of the single case in probability theory (e.g., Salmon, 1979). For example, assume that you want to estimate the probability that a person X has a heart attack before the age of 40. A routine procedure is to compute the proportion of heart attacks before the age of 40 within a reference class of persons matched to X on a number of relevant features (e.g., age, blood pressure). However, depending on the choice of features (i.e., the reference class) you arrive at different estimates. To get a good estimate, you prefer to enter as many relevant features as possible, but with all relevant features, the reference class may turn out to have a single member (X!).

An organism in an environment where the objects and situations are described by a large number of features is faced with an analogous problem. One possibility is to estimate the probabilities by computing a relative frequency as conditioned on a single feature (cue), as presumed by models of *cue-based inference* (e.g., Gigerenzer et al., 1991; Juslin, 1994). For example, in Table 1 the relative frequency of the name Larry conditional in presence of Cue 4 is 2/3 (i.e., the equivalent of a cue validity in Gigerenzer et al., 1991; Juslin, 1994). This, however, ignores the information provided by the unattended features. This is especially problematic in states of limited knowledge where it remains uncertain what the best cues are. A second possibility is to attend to all features and retrieve identical exemplars as implied by the Picky frequentist. This defines a reference class of high relevance, but the sample size is bound to be small or zero, again, especially in states of limited knowledge.

Similarity-graded probability provides a compromise between the demands for relevance and sufficient sample size: all exemplars are considered but similar exemplars receive a larger weight in the estimate (the weight is determined by the similarity parameter s). For example, assume that you are asked to assess your confidence that Indonesia has more than 70 million inhabitants. Perhaps, you only know the population of a small set of countries. The idea is that your judgment is informed by the distribution of values within this known set, but countries similar to Indonesia receive a larger weight in your judgment (see Nosofsky, 1998, for a more general discussion of exemplar models and optimal estimation).

In the next section, we verify that similarity-graded judgment is superior to both cue-based inference and a picky frequentist version of PROBEX for making probabilistic inferences, in

particular, in states of limited knowledge. Next, we present a modification of the context model, but the virtues detailed in this section are nonetheless preserved.

3. PROBEX—the algorithm

With PROBEX, the value of a continuous quantity or a subjective probability is estimated by rapid, sequential retrieval of exemplars from long-term memory. The activation of exemplars in memory is a parallel process, but the judgments arise from sequential processing of the retrieved exemplars. The search for exemplars is terminated once a clear-enough conception of the estimated quantity has been attained. For example, suppose that you are asked to estimate the population of Singapore. The probe Singapore activates similar cities in memory for which the population is somehow known (e.g., Djakarta, Shanghai, etc.), where the similarity parameter s determines the discriminability of the activation.

We now make the following terminological conventions explicit. We refer to entities in the environment as *objects* (e.g., the city of Djakarta), the memory representations of these objects as *exemplars*, and the object-name presented in a judgment task as a *probe*. The property to be estimated by PROBEX is referred to as the *criterion*. The criterion may either be a continuous quantity (e.g., population) or a binary index that signifies an event category (e.g., index 1 if the city population exceeds 4 millions and 0 otherwise). The probabilistic inference made by PROBEX may be elicited as a *point estimate* for a continuous variable, a *decision* between event categories, or a *subjective probability judgment*.

3.1. Environment and knowledge state

PROBEX exploits the structure of an environment as represented in the exemplars. The environment is modeled by a matrix, with D feature-dimensions and O objects. In our applications of PROBEX, we rely on values for D and O that are set *a priori* by the task. The *environment matrix* is projected into a *knowledge-state matrix* by filtering it with a *retrieval probability* p_r, the probability that a feature in the environment matrix is available from the knowledge-state matrix. The knowledge-state matrix has D feature-dimensions and K exemplars ($K \leq O$). With $p_r = 1$, all feature values are available from the knowledge-state matrix and PROBEX has perfect knowledge; with $p_r = .5$, half of the feature values are available from the knowledge-state matrix, and so on. The cells of the knowledge-state matrix not filled with “1” (“present”) or “0” (“absent”), are filled with “?” denoting an unknown value. A feature value is unknown either because it has never been encoded or because it has been forgotten.

Note that in this application of PROBEX to generic knowledge, exemplars do not correspond to memory traces of each single presentation with a stimulus, but to declarative knowledge-structures corresponding to objects in the environment (e.g., to Singapore). These crystallized knowledge structures retrieved from semantic memory may themselves have been formed by processes that involve abstraction across stimulus presentations. This interpretation departs from the interpretation in some previous exemplar or instance-based models (Logan, 1988; Nosofsky & Palmeri, 1997). Exemplars are coded in terms of binary features, except for

one continuous dimension (population) required by the task used to compare the ecological rationality of PROBEX with the other algorithms (Gigerenzer & Goldstein, 1996).

3.2. Activation in memory

When PROBEX is presented with a probe (an object-name) and the corresponding exemplar exists in memory, this exemplar is activated. If no exemplar that corresponds to the probe exists, a default vector for unknown probes is generated that only contains "?." In both cases, the (other) exemplars are activated as a function of their similarity to the probe and used to infer the criterion of the probe. Similarity is computed by the multiplicative similarity rule (Eq. (1) above), with one modification. If one feature value is unknown and the other known for feature dimension j, the feature index d_j is $\sqrt{s}$, interpreted as a half mismatch. If both feature values are unknown, the feature index d_j is 1, implying that they are similar in the sense of both being unknown. When knowledge of an object is positively correlated with its criterion (e.g., when we know more about cities with large population), the unknown–unknown similarities drive the *recognition principle* (Gigerenzer & Goldstein, 1996). For example, encountering an unknown city, it will be similar to other vaguely known cities, and if you happen to know the population of one such city it is likely to be small.[2]

Exemplars race to be retrieved (Logan, 1988; Nosofsky & Palmeri, 1997). The race is repeated, but the winners of a race do not participate in subsequent races. A stopping-rule decides when the sampling is terminated and a judgment is made. Exemplars $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_N$ are thus retrieved one-by-one from an initial set K_1 without replacement, $(i < n) \Rightarrow \bar{x}_i \notin K_n$. The probability that any exemplar $\bar{y}$ is sampled at iteration n is

$$p_n(\bar{x}_n = \bar{y}) = \frac{S(\bar{t}, \bar{y})}{\sum_{\forall \bar{z}, \bar{z} \in K_n} S(\bar{t}, \bar{z})}, \quad \forall \bar{y}\, (\bar{y} \in K_n). \tag{5}$$

The rationale is that it makes little sense to retrieve, for example, the population of Shanghai repeatedly, when you try to estimate the population of Singapore. The summation in the denominator is thus performed across the exemplars not yet sampled. A response is generated at iteration N, where the stopping rule specified below terminates the sampling. N is a random variable, the distribution of which predicts empirical response times. Note also that *all* algorithms rely on retrieval and thus need some equivalent to Eq. (5), regardless of whether it is cues (Gigerenzer & Goldstein, 1996) or exemplars that are retrieved.

3.3. Judgment process

3.3.1. Point estimation

To estimate the criterion $c(\bar{t})$ of probe $\bar{t}$, the criteria $c(\bar{x}_i)$ of retrieved exemplars $\bar{x}_i$ are considered. The estimate of the criterion $c'(\bar{t}, n)$ at iteration n is,

$$c'(\bar{t}, n) = \frac{\sum_{i=1}^{n} S(\bar{t}, \bar{x}_i) c(\bar{x}_i)}{\sum_{i=1}^{n} S(\bar{t}, \bar{x}_i)}, \tag{6}$$

a weighted average of the retrieved criterion values, where the similarities of the exemplars are the weights. The final estimate is $c'(\bar{t}) = c'(\bar{t}, N)$, where N is the first iteration where the

conditions for the stopping rule are satisfied. The final estimate $c'(\bar{t})$ is a best guess about the criterion value of probe $\bar{t}$ (e.g., the population of Singapore). Eq. (6) is merely a continuous version of Eq. (2) for the context model. Thus, if we define the criterion value as 1 if an exemplar belongs to Category A and 0 otherwise, Eq. (6) is equivalent to Eq. (2) (i.e., the sequential sampling aside). Similar extensions have been applied to perceptual classification by Andersson and Fincham (1996), function learning by DeLosh et al. (1997), and social prediction by Smith and Zarate (1992), but without sequential sampling.

3.3.2. Decisions

The original context model over-predicts the probabilistic responding in data, in particular, in individual-participant data (e.g., Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). Eq. (2) implies that the classification probabilities should match the ratio of the summed similarities (akin to so-called "probability matching"), but participants tend deterministically to prefer the category favored by the ratio (maximizing). That is, if Eq. (2) exceeds .5 almost all the participants' decisions favor Category A, if it is below .5 almost all decisions favor Category B. One way to modify Eq. (2) is by raising the summed similarities to a power θ, where θ is a free parameter determined by data (Maddox & Ashby, 1993; McKinley & Nosofsky, 1995). A problem with this formalization is that, although it improves the quantitative fit to data, the psychological rationale for it remains obscure.

PROBEX provides a more direct solution to this problem: exemplar retrieval is probabilistic but the decision rule is deterministic (see Nosofsky & Palmeri, 1997, for a similar approach). With PROBEX, Categories A or B may be predefined by a feature value present in the input vectors of the exemplars, as in standard classification experiments. Alternatively, the category is defined *post hoc* by the task instructions provided at retrieval. For example, in the application to data below, two Categories A and B are defined by reference to a cut-off d on the continuous criterion stated in the task. PROBEX assigns exemplar $\bar{x}_i$ to Category A if its criterion $c(\bar{x}_i)$ exceeds d and to Category B otherwise. Specifically, the population of a city is stored in the exemplar. The task instructions require a judgment of the probability of the "*post hoc* event" that a city has more than 180,000 inhabitants. In this case, the *predicted proportion* $P(\bar{t} \in \mathrm{A})$ *of decisions* that probe $\bar{t}$ is in Category A:

$$P(\bar{t} \in \mathrm{A}) = p(c'(\bar{t}) > d). \tag{7}$$

Prosaically, if a person estimates that the probe has a criterion above the decision criterion d, he or she assigns it to the category of objects with criterion values above the decision criterion (Category A). Eq. (7) applies also to standard classification designs with predefined categories that are not defined in terms of a continuous criterion variable. In this case, $d = .5$, and the criterion is simply 1 for Category A exemplars and 0 otherwise. It is easily shown that this is equivalent to a deterministic version of Eq. (2) for the context model.

Despite the deterministic decision rule, PROBEX produces matching in aggregated data. Because the sampling of exemplars is probabilistic, the decisions differ from trial to trial even if the decision rule is deterministic. Moreover, the matching (or variability) co-varies with the task difficulty, as defined by the difference between $c(\bar{t})$ and d. For difficult probes with a small difference between $c(\bar{t})$ and d, there is more variability. Thus, for example, in most

knowledge states, PROBEX produces more variability for the decision whether Helsinki has more or less than 500,000 inhabitants than for the decision whether Hong Kong has more or less than 500,000 inhabitants. This mimics the relationship between predictability (difficulty) and consistency in multiple-cue judgment research (Brehmer, 1994).

3.3.3. Subjective probability judgment

With minor modifications, Eq. (2) of the context model applies also to subjective probability judgment. Let event outcome index $e(\bar{x}_i \in \mathrm{A})$ be 1 for exemplars that belong to Category A and 0 otherwise. The predicted *judgment* $P(\mathrm{A})$ *of the probability* that probe $\bar{t}$ belongs to Category A after the terminal iteration N is,

$$P(\mathrm{A}) = \frac{(\phi/M) + \sum_{i=1}^{N} S(\bar{t}, \bar{x}_i) e(\bar{x}_i \in \mathrm{A})}{\phi + \sum_{i=1}^{N} S(\bar{t}, \bar{x}_i)} + \varepsilon, \tag{8}$$

where ϕ is a free parameter in the interval $[0, +\infty]$ for *dampening*, and M is the number of exclusive events into which the probability space is partitioned (e.g., with two alternative events, $M = 2$). The parameter ϕ dampens the effect of the retrieved exemplars when the sample size N is small. For example, without a dampening parameter, the probability is always 1 or 0 when the algorithm terminates after the first exemplar. The dampening implies a more modest estimate of $((\phi/M) + S(\bar{t}, \bar{x}_i))/(\phi + S(\bar{t}, \bar{x}_i))$ or $(\phi/M)/(\phi + S(\bar{t}, \bar{x}_i))$. When no exemplars are retrieved, the probability is assessed to be $1/M$ corresponding to a uniform "prior" probability (e.g., with $M = 2$, the "prior" probability is .5).

In the general case ($0 < s < 1$), Eq. (8) is a similarity-graded probability. However, in contrast to Eq. (2) of the context model, Eq. (8) resembles an optimal Bayesian estimate of a relative frequency (proportion). The dampening essentially plays the role of the α and β parameters of the β-distribution in Bayesian estimation (with the additional constraint that $\alpha = \beta$). Although, not part of a process model, Anderson (1990) introduced a dampening in his *rational model* (see also Nosofsky, Kruschke, & McKinley, 1992). ε is a normally and independently distributed response error with mean 0 in the use of the overt probability scale. The judgments from Eq. (8) are truncated to fall in the interval $[0, 1]$: values larger than 1 are assigned probability 1, values below 0 are assigned probability 0 (Juslin et al., 1997, 1999).[3] The response error is controlled by the *response error parameter* σ_r^2.

3.3.4. Stopping rule

The rule for terminating exemplar retrieval may differ depending on the task (e.g., whether a point-estimate or a categorization is the focal concern). In the applications presented below, the task revolves around a continuous criterion dimension (i.e., city-population). Therefore, we concentrate on a simple stopping rule appropriate to this task. Sampling is terminated at the first iteration N where this condition is satisfied as

$$|c'(\bar{t}, n) - c'(\bar{t}, n-1)| < k|c'(\bar{t}, n)|. \tag{9}$$

The free parameter k decides the sensitivity of the stopping rule. Intuitively, one can think of this rule as a way of judging when the change in the point estimate from $c'(\bar{t}, n-1)$ to $c'(\bar{t}, n)$ is too small to merit continued sampling. Although the stopping rule is error-prone because successive point estimates are sometimes equal by chance, it minimizes

the need to store intermediate steps in the process. Parameter *k* determines the extent of retrieval: for small *k* PROBEX samples extensively, but for large *k* it terminates after retrieval of a few exemplars. Parameter *k* is sensitive to cost–benefit considerations and affords implementation of an exemplar model in line with the notion of bounded rationality.

One important aspect of PROBEX is that it belongs to the class of *lazy algorithms* that has proven useful in artificial intelligence due to their flexibility (Aha, 1997). This latent property of exemplar models means that they need not rely on *pre-computed knowledge*. (Pre-computed knowledge may nevertheless enter to the extent that an exemplar model modifies its attention weights to a specific task, see, e.g., Nosofsky, 1986.) All computations by PROBEX are performed at the time of the judgment. Algorithms extensively premised on pre-computed knowledge (e.g., cue validities) soon become untenable. To attain flexibility they require ominous foresight for future task demands or enormous amounts of pre-computed knowledge.

In a previous section, we proposed that similarity-graded judgment afford an advantage in states of limited knowledge. In this section, we propose that PROBEX furnishes such an algorithm that applies similarity-graded judgment to generic knowledge, making psychologically plausible demands on storage, retrieval and computation. To substantiate these claims, we turn next to an analysis of the ecological rationality of PROBEX.

4. The ecological rationality of PROBEX

Gigerenzer and Goldstein (1996) demonstrated that simple heuristics as *Take-The-Best* (TTB) are able to compete evenly or even outperform more complex algorithms like linear multiple regression when applied to a real environment. The task in the original study on ecological rationality was the German city-population task, where the algorithms are fed with pair comparisons such as "Which city has the larger population: Heidelberg or Erlangen?" TTB searches the cue values for each city in a particular order (defined below) and makes a decision for the first pair of cues that differentiate between the cities. Comparing binary numbers is a good analogy: 00101001 is larger than 00100110, because the fifth digit differentiates between the numbers. In this case the decision is always correct.

Knowledge of the environment consists of nine binary cue values for each of 83 German cities, for example, whether the city is state-capital or not, whether it has a university or not. The *ecological validity* of a cue is defined by the relative frequency with which it selects the correct answer when applied to all pair-wise comparisons between the 83 German cities. For example, consistently choosing a city with a university over a city with no university leads to a relative frequency of correct decisions equal to .71. TTB relies on the *first* most valid cue that is applicable to a question. Multiple regression integrates all nine cues into an optimal population estimate for each of two cities and decides on the one with the higher estimate. The idea is that, where as multiple regression embodies the ideals of *classical rationality* implying unlimited time, knowledge and computational power, TTB provides a psychologically plausible alternative satisfying *bounded rationality* (Simon, 1990).

The results presented by Gigerenzer and Goldstein (1996) support two conclusions. First, the rationality of an algorithm cannot be properly evaluated without careful attention to the environment that provides the behavioral support for the algorithm. Second, considering the impressive accuracy and the low computational cost of TTB, it provides a viable alternative to more complex algorithms: It is "fast and frugal."

4.1. Varieties of frugality

One crucial aspect of the argument presented by Gigerenzer and Goldstein (1996) concerns the frugality of the algorithms. In the following, we broaden the discussion of frugality in Gigerenzer and Goldstein (1996) by addressing three complementary aspects of frugality: demands on storage, computation, and knowledge access.

4.1.1. Storage

With PROBEX, there is no need for pre-computed knowledge, over and above storage of exemplars. While the extensive storage of exemplars is consistent with the huge storage capacity of long-term memory, the memory demands implied by exemplar models is sometimes raised as a concern (e.g., Nosofsky et al., 1994). While this concern is difficult to evaluate without a proper theory of the cost and other limitations on storage, it is mainly warranted in the case of exemplar models that presume *obligatory storage* of each object, or even each presentation of an object, such as Logan's (1988) instance-race model (see the discussion in Barsalou et al., 1998). This is not presumed by PROBEX: it only states that whatever exemplars have been stored provide input to similarity-based inference. As illustrated below, PROBEX is remarkably robust when few exemplars are stored.

The processing of situations or objects in one way or another is required by any algorithm, explicitly modeled or not. In regard to TTB and multiple regression we discern two possibilities. First: abstractions in the form of cue validities or beta-weights are continuously computed and updated as new objects are encountered with no memory storage of the objects. This is feasible when the task is known already when the objects in the environment are encountered. For example, a person knowing that he or she is later required to predict population from the nine cue values of German cities, may (somehow) compute the nine cue validities appropriate to this specific task and later apply the TTB algorithm.

When future task demands are difficult to foresee at the time of exposure with the objects, the alternative is to compute *all* contingencies between the variables that describe the objects (what might be referred to as "*obligatory computation*"). The German city-population task with 10 variables ($C = 10$) requires pre-computation of 90 cue validities for TTB and 90 beta-weights for multiple regression. Because the number of pre-computed abstractions is $C^2 - C$, the number of stored abstractions increases rapidly (e.g., 9,900 for $C = 100$, 999,000 for $C = 1{,}000$). Clearly, as a tool to make inferences for the multitude of unforeseen tasks that define our everyday environment, this is not a frugal solution in terms of storage.

A second possibility is that TTB and multiple regression store all exemplars and computation of abstractions is postponed to the time of the judgment, that is, to the time where the abstractions required by the specific task are known. As detailed next, this poses problems for the idea that TTB is a fast and frugal at the time of the judgment.

4.1.2. Computation

To compute the cue validity for each cue, TTB has to check every pair of cities that can be generated from the knowledge matrix. For each pair, it has to ascertain whether the cue discriminates between the cities, and if it does, if the cue selects the correct or the wrong answer. The number of checks is $.5(K(K-1))C$, a number ranging between nine checks per cue for training set-size 2 to 28,440 checks at training set-size 80. The computational complexity to derive the order of the cue validities grows with the square of the number of exemplars K (i.e., with CK^2). Regardless of whether this is done continuously as the objects are encountered as a matter of automatic processing of frequency, or at the time of the judgment, it amounts to extensive computation. (With multiple regression the computational complexity becomes even more daunting.) However, if all cue validities are pre-computed and ordered, TTB only has to check between 1 and C cues to make a decision. This is the frugality stressed by Gigerenzer and Goldstein (1996).

With PROBEX, the extent of computation is determined by the number of sampled exemplars, which in turn depends jointly on the parameter k and the difficulty of the task. In particular, the parameter k can be set to make PROBEX sample only a few exemplars to make a judgment, thereby instantiating a fast-and-frugal exemplar model. Note that the access to exemplars is a matter of *retrieval*—much as when TTB retrieves the cues. The difference is that TTB (and multiple regression) retrieves knowledge in the form of abstractions that have been pre-computed for very specific inferential purposes (i.e., cue validities).

4.1.3. Knowledge-access

The German city-environment is a matrix with 83 rows, one for each German city, and 10 columns, nine binary cues and one continuous dimension, population. States of limited knowledge can be modeled by knowledge of 2 and 80 cities. Note that at least 3 cities are always left for the test set, thus the maximum of 80 cities. TTB computes the cue validities, and multiple regression, the beta-weights from the set of all known cities. The cells in the matrix accessed by these algorithms is therefore 10 times the number of known cities, as indicated by the dotted line in Fig. 1.

In PROBEX, activation of exemplars is an automatic and parallel process that produces *retrieval* from memory, much as when a probe is assumed to elicit retrieval of cues with TTB or multiple linear regression. All algorithms presume retrieval, although these are not explicitly spelled out in the discussions of TTB or multiple linear regression (Gigerenzer et al., 1999). In PROBEX, the *judgment process*—corresponding to cue substitution with TTB and cue integration with multiple regression—is the integration of retrieved exemplars.

If PROBEX samples all of the exemplars ($k = 0$), it will access the same number of cells. Fig. 1, however, plots the number of cells accessed by PROBEX as a function of the number of known cities (training set size) for different stopping parameters k. The similarity parameter s in the execution of PROBEX was .1. (The other two parameters of PROBEX refer to overt probability assessments and have no effect on the decisions in the German city-population task.) The curves plotted in Fig. 1 are the results of applying the PROBEX algorithm outlined above to the German city-population task. Regardless of the number of known cities, the minimum number of cells accessed by PROBEX is 20. This occurs with very high values of the stopping parameter k. In this case, PROBEX retrieves only one exemplar for each probe in

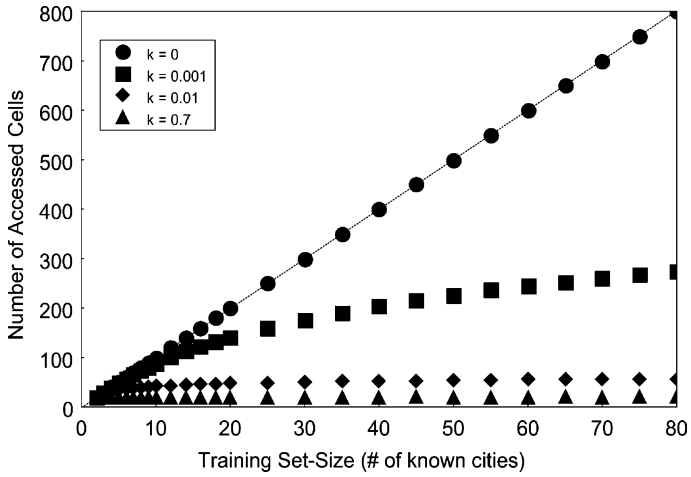


Fig. 1. The mean number of cells accessed in the declarative knowledge base for the German city-population task as a function of the number of known cities (the training set size). The dotted line represent the cells accessed by TTB and linear multiple regression. The solid lines are the number of cells accessed by PROBEX for different stopping parameters k.

the pair-comparison. The values of k plotted in Fig. 1 are 0, .0001, .01, .7; $k = .7$ is the value fitted to the data presented in the empirical section.

In general, PROBEX relies on a small fraction of the cells retrieved by the other algorithms. For example, at $k = .01$ and 40 known cities, PROBEX will only access 13% of the cells accessed by TTB and multiple regression. At $k = .7$ and 29 known cities, parameter values that are appropriate to empirical data below, PROBEX will access 8% of the cells accessed by the other algorithms. As the simulations presented below demonstrate, for appropriate values of the s parameter the accuracy of the decisions is largely unaffected by the number of sampled exemplars (as determined by k). Simulation 3 below demonstrates that when k is increased from 0 (exhaustive sampling) to .01, the accuracy in the binary discrimination task between German city-populations only decreases from .637 to .630, but the percentage of cells accessed decreases from 100% to only about 10%. Fig. 1 thus suggests that PROBEX maintains high accuracy while effectively minimizing the need to access knowledge. We propose that, considering the flexibility required in a complex environment with unforeseen demands, the frugality favored by evolution is likely to be that of a "lazy algorithm."

4.2. Accuracy and robustness

To get a benchmark it is natural to compare PROBEX with TTB and linear multiple regression in the German city-population task. Gigerenzer and Goldstein (1996) tested the algorithms by feeding them all the pair-wise comparisons between German cities with more than 100,000 inhabitants. A weakness of the test procedure was that the knowledge of the algorithms was assumed to consist of all cities (Persson, 1996). When the experience consists of a small sample, the sample will sometimes suggest the wrong cue-direction (e.g., when, due to sampling

error, only small cities with universities have been encountered, suggesting that university predicts small population). A better test (Czerlinsky, Gigerenzer, & Goldstein, 1999) is to split the set of German cities into one training and one test set, where ecological cue validities are estimated from the training set. This set is thus used to learn the parameters of the algorithms (e.g., beta weights) that are then tested on tasks from the test set. This cross-validation is a true test of the robustness of an algorithm and detects over-fitting to the training set. (See Persson & Juslin, 2000, on the crucial importance of detecting the correct cue-directions in the German city-population task, cf. Dawes, 1982.)

We also tested the algorithms when the number of cities in the training set was very low. From an evolutionary perspective, it is important that a decision algorithm is good when information is low, because the decision-maker has to survive also as a "neophyte." Moreover, the performance of any useful algorithm soon converges on the limit imposed by the knowledge state (i.e., the nine cues in the German city-population task). Thus, the small training sets most clearly distinguish between good and poor algorithms.

4.2.1. Simulation 1: Accuracy of binary decisions

PROBEX competes with TTB and linear multiple regression. The size K of the training set (i.e., the cities known to the algorithms) is varied between 2 and 80 cities. To give regression a fair chance also for the smaller training sets (e.g., where the number of observations may be smaller than the number of beta-weights to be estimated), we relied on ridge-regression[4] (see Garg, 1984, for an introduction). We also provide the data for ordinary regression. In regard to frugality, it is of interest to investigate the accuracy of a "minimalist-version" of PROBEX where only a single exemplar is retrieved for each of the two probes in the two-alternative task (i.e., instantiating a "nearest neighbor algorithm"). Minimalist PROBEX implies no information integration, only the similarity-based retrieval of a single pair of exemplars. Hence, while the other algorithms rely on all known objects, Minimalist PROBEX constrains the computation and retrieval to an absolute minimum.

4.2.1.1. Method. The proportion of correct decisions among the pair-comparisons of the test set was the dependent variable. For each training set-size (2–80), 1,000 participants were simulated. For each simulated participant, the German cities were randomly partitioned into new and independent training and test sets. PROBEX was executed with $s = .5$ (a similarity-graded probability) and $k = 0$ (exhaustive sampling of the training set). The other algorithms estimated task-specific parameters from the training set (e.g., beta-weights appropriate for the German city-populations). The remaining cities defined the test set (size 3–81). The data for the 83 German cities was collected from the Appendix of Gigerenzer and Goldstein (1996).

4.2.1.2. Results and discussion. The highest proportion of correct decisions was achieved by PROBEX (.69), closely followed by both TTB and ridge regression (.68 for both). The lowest accuracy was attained by multiple regression (.57), which was even beaten by Minimalist PROBEX (.61).[5] Note that the test procedure used here to provide a stringent test of the algorithms puts PROBEX at a particular disadvantage, as compared to a real-life application. In realistic applications, it is likely that objects from both the training and test sets appear in future judgment tasks. Because the training objects are stored as exemplars in PROBEX, the

correct criterion values for these objects can be retrieved directly. In this case, the accuracy of PROBEX increases. The other algorithms attain the same beneficial effect of testing against both training and test sets only by storing *both* all exemplars and abstractions (e.g., beta-weights).

The results are presented in Fig. 2A. Because the algorithms converge on similar asymptotes at high training set sizes, the *x*-axes in Fig. 2 are in logarithms, thus emphasizing the range were "the action" is. When PROBEX samples all exemplars of the training set—as is always done by TTB and regression—it dominates over its competitors regardless of the training set size. The performance of multiple regression is extremely poor, once that the parameters have to be estimated from the training set (i.e., rather than calculated from the entire set of objects and provided *a priori* to the algorithm). The explanation is that a complex algorithm like multiple regression is particularly susceptible to the problem of over-fitting. Thus, while regression may provide a good account of the training set, the beta-weights from the training set fail to generalize appropriately to the test set. Over-fitting is more than a technical problem—it is a real and profound challenge for any cognitive algorithm. This poor performance contrasts sharply with the robustness of simpler algorithms, like PROBEX and TTB. On the smallest training sets, ridge regression performs on par with PROBEX, but this is due to the ridge constant chosen to fit very small samples.

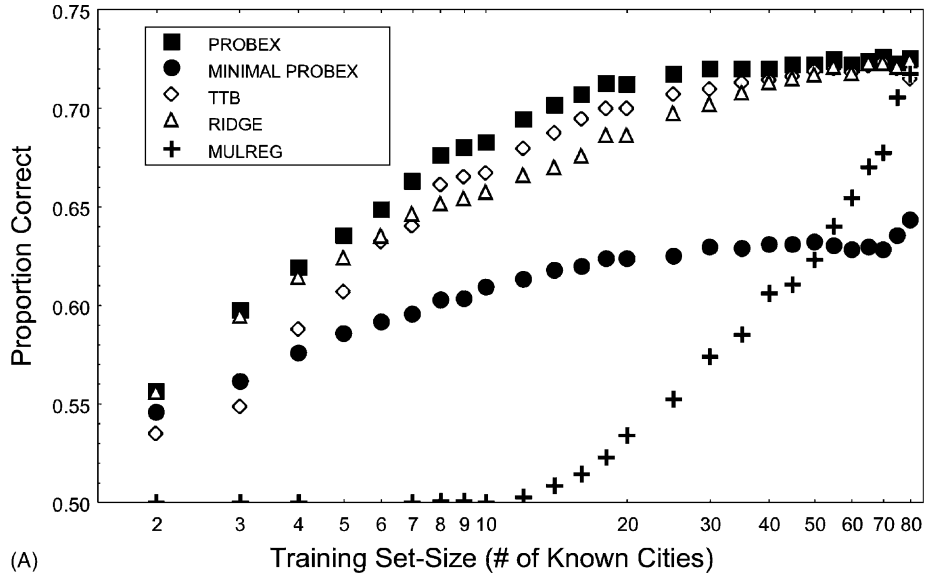
Minimalist PROBEX performs poorer than its competitors, but nonetheless surprisingly well considering the minimal computational cost that is invested in the judgment. Arguably, if we take into account *both* the computations performed prior to and at the time of judgment, it is difficult to come up with a more frugal algorithm. A complex algorithm, like multiple regression is only better than Minimalist PROBEX when the training set exceeds 55 cities. The Minimalist PROBEX relies on a similarity parameter $s = .1$, a similarity-graded probability. Because only one exemplar is sampled, it is important that this exemplar is similar to the probe and *s* should thus be on the "picky" side (i.e., low).

The weak spot of TTB is inferences from few training exemplars. If few objects are known, there is not enough information to identify the correct order of the cue validities and it is better to consider all cues and integrate all of this information into a similarity-graded judgment. If experience is allowed to shape the rank-order of cue validities list to perfection, however, TTB is an efficient algorithm, at least for often-repeated tasks.

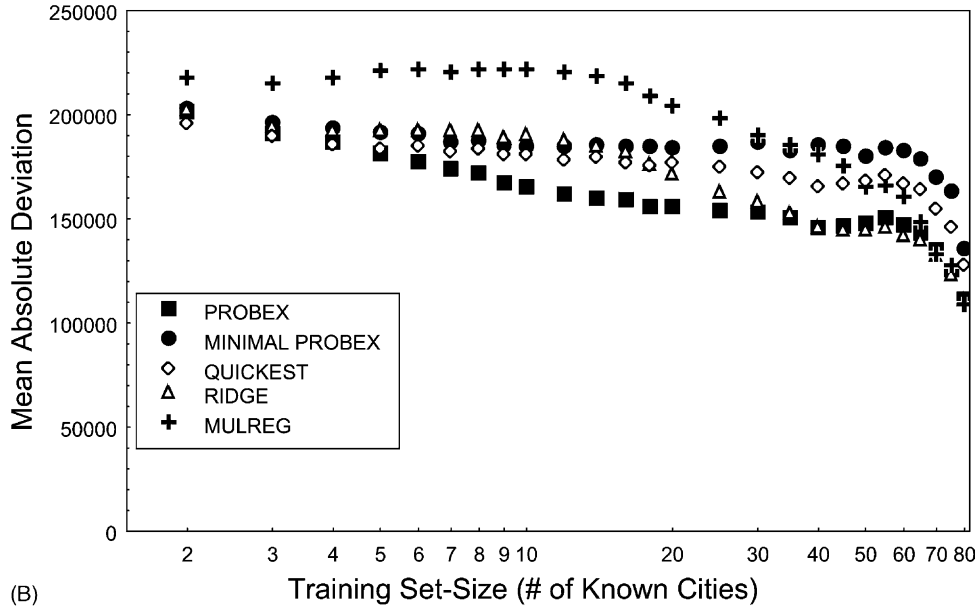
Simulation 1 illustrates two aspects of the versatility of lazy algorithms such as PROBEX. First, when the algorithms benefit from the complete sample of *K* known objects, PROBEX dominates over its competitors, in particular in states of limited knowledge ($K < 30$). Second, Minimalist PROBEX performs better than chance *without any* pre-computed knowledge and *no* information integration at the time of the judgment, even outperforming multiple regression for training sets containing less than 55 cities.

4.2.2. Simulation 2: Precision of point estimates

In Simulation 1, the precision of the point estimates was obscured by the binary choice task. Because PROBEX and linear multiple regression provide explicit point estimates the precision of these estimates can be more carefully investigated. TTB does not apply to this task and was therefore replaced by a companion fast-and-frugal algorithm appropriate to the point-estimation task, *QUICKEST* (Hertwig, Hoffrage, & Martignon, 1999).



(A)



(B)

Fig. 2. (A) Proportion correct for PROBEX, Minimalist PROBEX that only samples a single exemplar, Take-The-Best (TTB), linear ridge regression (RIDGE), and linear multiple regression (MULREG) in the German city-population task, plotted as a function of the logarithm of the number of known objects (German cities). (B) Mean absolute deviations between predicted and actual German city-populations for the same algorithms, except that TTB has been replaced by the algorithm QUICKEST (see the text for details).

QUICKEST is appropriate to skewed distributions like that of the German city-populations, where most cities have small populations. For each cue, the mean population for cities with negative cue values is computed (negative cue values are those that go with small populations, e.g., not being a state capital). Cues are rank-ordered from the cue with the lowest mean given a negative cue value to the cue with the highest mean given a negative cue value.[6] To estimate population the algorithm starts by checking if a city has the negative cue value that is first in this rank-order, then the next, and so on until a match is encountered. As soon as a match is encountered, the conditional mean associated with this cue value is reported as the estimate. Given the skew of the city-population distribution with mostly small cities, this algorithm is frugal in the sense of minimizing the number of cues that have to be accessed. That is, because most cities are small, the algorithm often stops for the first negative cue-value in the rank-order. As with TTB, QUICKEST is frugal at the time of judgment *per se*, but needs a considerable amount of pre-computation to be executed.

4.2.2.1. Method. The method was similar to the one used in Simulation 1, with the qualification that the mean absolute deviation of the estimates was the object of analysis.

4.2.2.2. Result and discussion. The overall mean absolute deviation was 158,528 for PROBEX, 166,932 for ridge regression, 173,001 for QUICKEST, 183,247 for Minimalist PROBEX, and 191,885 for multiple regression. PROBEX is overall the most robust algorithm, although QUICKEST is as good with 2–5 exemplars in the training set and all algorithms converge to the same performance at large training sets (see Fig. 2B). Ridge regression is best with much information. Again, ordinary multiple linear regression shows an extreme lack of robustness and performs poorly except when almost all objects are included in the training set. When the training set is less than 20 cities, Minimalist PROBEX is surprisingly close to algorithms that take advantage of all objects in the training set.

4.2.3. Simulation 3: The role of similarity

In the judgment literature, frequencies have generally been affiliated with rationality and similarity with irrationality (cognitive biases). Depending on the similarity parameter s of PROBEX, it can implement two kinds of frequentistic algorithms. The sloppy frequentist ($s = 1$) enters all exemplars alike and reports the base-rate, and the picky frequentist ($s = 0$) enters only identical exemplars into the judgment. In the general case ($0 < s < 1$), PROBEX responds with a similarity-graded probability. We address two questions: (a) Does an algorithm that exploits the similarity-structure of real environments confer an advantage over purely frequentistic algorithms? (b) What parameters of PROBEX provide the best balance between high accuracy, robustness, and low computational cost?

4.2.3.1. Method. The same procedure was used as in Simulation 1, but the parameter s was tested in the range of .1 to 1 and the size of the training set was 2, 5, 10 and 40 cities. The following values of the stopping rule parameter k were investigated: 0, .01, and .1. When $k = 0$, the sampling of exemplars will not stop until the entire training set of K exemplars has been exhausted (as all algorithms except Minimalist PROBEX did in Simulation 1).

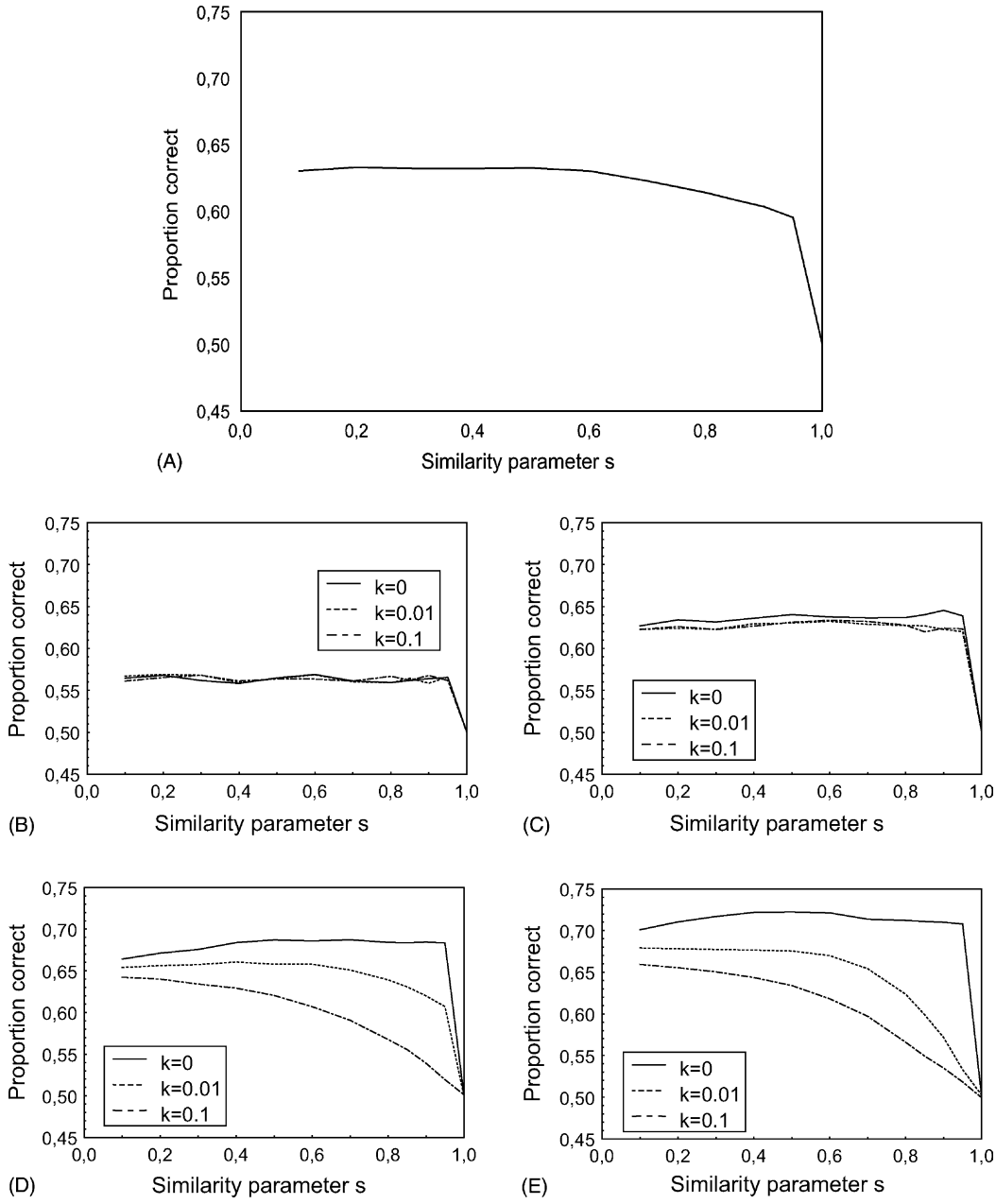


Fig. 3. (A) The overall proportion correct for PROBEX as a function of the similarity parameter s. (B) The proportion correct when the training set contains two cities. (C) The proportion correct when the training set contains five cities. (D) The proportion correct when the training set contains 10 cities. (E) The proportion correct when the training set contains 40 cities. ($k = 0$ amounts to sampling the entire training set of exemplars, $s = 0$ is the picky frequentist, and $s = 1$ is the sloppy frequentist.)

4.2.3.2. Results and discussion. We evaluate the effect of the similarity parameter *s* in two respects: overall accuracy and robustness. First, consider the intermediate levels of *s*, $s = .1, .2, \ldots, .9$. The main effect across the three levels of *k* (determining the number of sampled exemplars) and the four training set sizes *K* indicates a higher overall accuracy for *s*-values below .5 (see Fig. 3A). In this range the overall accuracy is approximately .63. These results suggest a slight advantage for similarity-graded probability on the "picky-side."

The results for each level of parameter *k* and for each size of the training set *K* allow us to understand the nature of this advantage (see Fig. 3B–F). As expected, the accuracy increases as the training set size *K* increases from 2 objects (Fig. 3A) to 40 objects (Fig. 3F). Overall, accuracy also increases with more extensive sampling (i.e., for lower *k*-values). More interestingly, however, there is an interaction between these two factors. At low values of *s*, accuracy is little affected by the number of exemplars sampled (i.e., by the parameter *k*) and performance is reasonably close to the accuracy at complete sampling of the training set ($k = 0$). Thus, lower *s*-values are superior in regard to robustness.

The accuracy of PROBEX with the extreme values of *s* that implements purely frequentistic algorithms is considerably poorer. When *s* is 1, the sloppy frequentist, all exemplars are retrieved for both of the two compared cities and the two estimates become trivially identical. This only allows random performance (proportion correct .5). A picky frequentist ($s = 0$), on the other hand, finds few or no identical exemplars to retrieve, in particular at the smaller training sets. Given the low number of identical cue-patterns in regard to the nine cue-dimensions for the German city-populations, this only allows for a performance marginally higher than expected by chance. Notably, only similarity-graded probability provides both sufficient relevance of the retrieved exemplars and a sufficient number of exemplars to yield useful inference from limited knowledge.

5. The psychological plausibility of PROBEX

In Juslin, Nilsson, and Olsson (2001), PROBEX was tested against three alternative models of probability judgment in a category learning task: the *representativeness heuristic* interpreted either as similarity to category prototype or as relative likelihood (Tversky & Kahneman, 1983), and *cue-based relative frequency* (Björkman, 1994; Gigerenzer et al., 1991; Juslin, 1994). (Because of its vagueness, representativeness was thus entered in two versions.) The participants task was to predict if the stock value of 15 fictive companies would rise or fall in the next year. Each company was described by four binary features. With outcome feedback, the participants were trained to make rise/fall decisions in four training blocks á 60 trials each. In a test phase that followed after each training block, they assessed the *probability* that the stock value of each company would rise in the next year. The category structure was created to disentangle predictions by the four models (see Juslin et al., 2001).

Fig. 4 presents the results of fitting two-parameter versions of each model to the probability judgments made in the four test blocks. PROBEX provides a superior fit to data in all four test blocks, both in terms of the root mean square deviation (RMSD) between the predictions and data (Fig. 4A) and variance r^2 accounted for (Fig. 4B). In the last test block, PROBEX

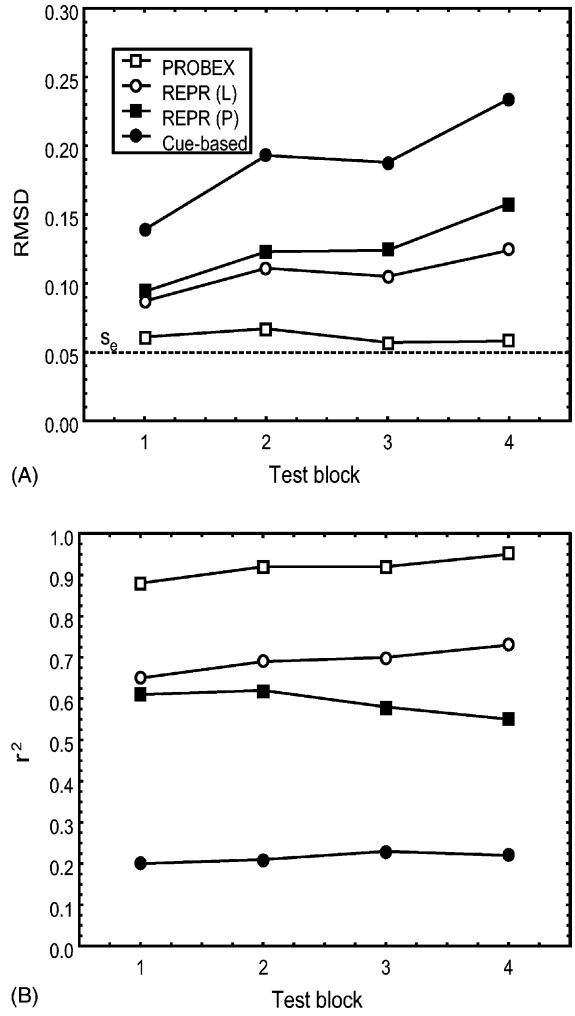


Fig. 4. The fit of PROBEX, representativeness as relative likelihood (REPR(L)). Representativeness as similarity to category prototype (REPR(P)), and cue-based relative frequency (cue-based) to the probability judgments in a category learning experiment. (A) RMSD between predictions and data as a function of test block (root mean square error in data, $s_e = .05$). (B) Coefficients of determination r^2 as a function of test block. Adapted from Juslin et al. (2001).

accounts for 95% of the variance in data with an RMSD of .058 (to compare with a root mean square error in data of .05). It seems clear that the judgments are responsive to both similarity and frequency in the manner modeled by PROBEX (see Juslin et al., 2001, for details and Sieck & Yates, 2001, for additional support for exemplar models).

In this article, we complement these results by applying PROBEX to a generic knowledge task of the sort commonly used in research on probability judgment. The prime questions were: (1) Is PROBEX a viable description of the inferences in a generic knowledge task? (2) If so, will the fitted parameters suggest a fast-and-frugal exemplar model?

In the experiment, we investigated the overall fit of PROBEX to the point estimates, binary decisions, subjective probability judgments, and response times by humans. The task was similar to the generic knowledge task used in the analysis of ecological rationality. The point estimates thus concerned "best guesses" about the population of German cities ("What is the population of Bonn?") and the decision tasks concerned whether presented German cities have populations above or below 180,000 inhabitants. In addition, the participants made probability judgments with three different judgment formats.

In research on subjective probability judgment, one of the most pervasive effects, in terms of the magnitude, is *format dependence* (Juslin et al., 1999; Klayman, Soll, Gonzales-Vallejo, & Barlas, 1999). In *calibration* studies, subjective probabilities are compared to "objective" probabilities, often by plotting relative frequencies against subjective probability categories in calibration graphs. If the subjective probabilities are realized in terms of the corresponding relative frequencies, the participants are calibrated (e.g., if 70% of the events assigned probability .7 occur). Particular attention has been paid to the assessment of too extreme subjective probabilities, so-called *overconfidence bias* (e.g., Erev, Wallsten, & Budescu, 1994; Gigerenzer et al., 1991; Juslin, Winman, & Olsson, 2000).

Format-dependence entails that with a given task-content there is underconfidence when participants assess a probability on a scale between .5 and 1, and yet overconfidence when the *same participants assess the same task content* on a scale between 0 and 1 (Juslin et al., 1997). Moreover, when the same task content is approached with interval estimation there is often extreme overconfidence (Juslin et al., 1999). In the experiment presented below, the three formats (which we describe in detail afterwards) were provided in a within-subjects design. The prime concern was whether PROBEX could reproduce format-dependence.

With the *half-range format*, the participants decided on one of two alternative answers and assessed confidence in this decision as a probability between .5 and 1. For example:

The population of Hannover exceeds 180,000 inhabitants.

(a) True			(b) False		
50%	60%	70%	80%	90%	100%
Random					*Certain*

With this format, over/underconfidence is measured by the difference between the mean probability assigned to the chosen answer and the proportion correct, where a positive difference indicates overconfidence bias. For example, if the participants are 90% certain on average, but only have 80% correct answers, there is 10% overconfidence. The *full-range format* requires an assessment of the probability that a statement is true on a scale from 0 to 1.

The population of Hannover exceeds 180,000 inhabitants.

What is the probability that this statement is true?

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Certainly false										*Certainly true*

Clearly, if the participant responded with "True" and "90%" confidence in the half-range task, he or she should assess the full-range probability to "90%." In full-range, overconfidence

is evidenced by too low proportions of true statements in high subjective probability categories and too high proportions of true statements in low subjective probability categories (i.e., by too extreme subjective probabilities). Overconfidence is indexed by transforming the full-range probabilities into half-range decisions and half-range probability judgments. Thus, full-range probabilities above .5 are taken to imply decisions in favor of the statement with the assigned probability as the half-range probability. Full-range probabilities below .5 imply a decision that the statement is false with the half-range probability equal to one minus the stated full-range probability (e.g., a full-range probability of .3 is interpreted as implying a decision "False" with half-range probability .7). After this transformation, over/underconfidence is computed in the same manner as for the half-range format.

Interval estimation requires assessment of the upper and lower bound of an interval intended to include the true value of a quantity with a pre-stated probability. For example: "Provide the smallest interval within which you are 80% certain that the true population of Hannover lies. The population of Hannover lies between _____ and _____ inhabitants with 80% confidence (or probability .8)."

Less transparently, perhaps, a participant responding with the decision "True" and "90%" confidence in the above half-range task should assess 180,000 inhabitants as the upper limit of his or her .8 interval. Overconfidence is measured by the difference between the pre-stated probability and the observed proportion of true values that fall within the interval. For example, if 50% of the true values fall within the 80% interval, overconfidence is 30%.

From a formal standpoint, these three formats are merely different ways of eliciting the same subjective probability distribution. The phenomenon of format dependence, however, entails that they lead to contradictory results (see Juslin et al., 1999, for a discussion). When PROBEX was fitted to data, we thus required it, not only to generate appropriate point estimates, decisions, probability distributions and response times, but also to map the inconsistent pattern of over- and underconfidence across these three formats.

To Swedish undergraduates many German cities are rather unfamiliar. Therefore, we expected them often to rely on what Gigerenzer and Goldstein (1996) refer to as the recognition principle. When applied to the choice task with pairs of German cities, this principle implies that when one but not the other of the cities is recognized, the participant guesses that the recognized city has the higher population. To implement all three assessment formats in a comparable manner, the tasks used in the experiment reported below concern the population of a single German city. In this task, we interpret the recognition principle to imply that known cities have larger populations on average than unknown cities. We also extend the recognition principle to cover limited knowledge of cues. If the probability of knowing a city is low, the probability of knowing each associated cue-value is low also.

Note that in PROBEX cue-based inference and the recognition principle are implemented by the same process. For example, Swedish participants might find Leverkusen to be more similar to Höcklenspiel than to Berlin, even though little, or perhaps nothing, is known about Leverkusen and Höcklenspiel. Unfamiliar cities have many unknown features ("?") and—as detailed above—these provide just another feature in the similarity analysis. Thus, if the participant knows the population of one, otherwise unfamiliar, German city—presumably, a small city due to the environmental correlation between recognition and population-size—he or she will guess that Leverkusen has a small population. To model how many, and which,

German cities were known to the participants, they also indicated which of the 83 German cities used in the study they recognized.

Finally, PROBEX does not allow analytic procedures to find optimal parameter estimates. Furthermore, the dependent variables are defined in terms of different metrics (i.e., the point estimates, decisions, and subjective probabilities). There is no obvious scheme for weighting together prediction errors into one measure of goodness-of-fit. We relied on a grid-search with the criterion of finding the parameters that simultaneously produced the best relative fit to data for each dependent variable (see Appendix A for definition).

In the experiment, we therefore concentrate on the following issues: (a) Will PROBEX and humans make similar point estimates? (b) Will the same items appear "misleading" to both PROBEX and humans? (c) Will PROBEX and humans produce the same calibration and display the format-dependence effect? (d) Will the pattern of response times by humans be well predicted by PROBEX? Finally, conditionally on affirmative (or, at least encouraging) answers to these questions. (e) Will the best-fitting parameters suggest that the participants approach the task with a fast-and-frugal exemplar algorithm?

5.1. Methods

5.1.1. Participants

Forty-one undergraduate Psychology students at the University of Uppsala participated in the experiment. There were 22 male and 19 female participants, with an average age of 25 years. The participants volunteered in order to get a course credit.

5.1.2. Materials

The 83 cities listed in the Appendix of Gigerenzer and Goldstein (1996) were used as the objects of inference (i.e., all German cities which then had more than a 100,000 inhabitants). A computer randomly selected the 40 cities presented to the participants from this pool of 83 cities. The task was to estimate the population of the presented German city, and to compare this population to the median city population within the overall set of 83 cities (i.e., 180,000 inhabitants) in terms of three assessment formats.

For each half-range item, participants decided whether a statement is true or false, and thereafter assessed confidence as a probability on a scale between "50%" (random choice, or pure guessing) and "100%" (perfectly certain of having chosen the correct answer). For each full-range item, the participant assessed the probability that the corresponding statement is true on a scale between "0%" (certainly false) to "100%" (certainly true). In the half-range and full-range items, the statements implied that the German city has a population above 180,000 inhabitants, as illustrated in the introduction to this experiment. For interval estimation, the participants assessed the smallest intervals that included the true value of the population of the German city with probability 1.0 (100%). The task items were presented one by one on a computer screen and the computer also controlled the order of presentation.

5.1.3. Design and procedure

The design of the experiment consisted of three within-subjects conditions: (a) items of the half-range format, (b) items of the full-range format, and (c) items of the interval

estimation format. The items in each condition were presented in blocks that contained the same set of 40 randomly selected items (German cities), presented in the same order. The order in which the three blocks (i.e., conditions) were presented to the participants was counter-balanced across participants. In each condition, the participants also made a point estimate of the number of inhabitants in the city for which the probability was assessed. For every other item in each condition this was made before the probability was assessed, for the other items it was made after the probability assessments. Finally, a paper form presenting all 83 cities in alphabetical order was given to the participants, asking them to indicate which of the cities they recognized *before* the experiment by underlining their names.

Each participant was run individually and each session took approximately 1 h. With the half-range format, it was explained in the instruction that .50 (50%) meant "random" and that 1.0 (100%) meant "certain." Participants also received a brief written tutorial on the concept of calibration, and it was stressed that for each particular item, they should select the subjective probability value that best reflected their degree of confidence for this particular item. In the full-range instruction it was explained that probability 0 (0%) meant that the statement was "certainly false." and that probability 1.0 (100%) meant that the statement was "certainly true." Otherwise, the instructions were the same as in the half-range condition. In the instruction for interval-estimation it was explained that the 1.0 probability interval meant that there should be no doubt whatsoever that the true value was inside the interval. In other respects, the instructions were the same as in the other two conditions.

5.2. Results

5.2.1. Fitting PROBEX to data

PROBEX was applied to a data matrix, with nine cue-dimensions, one target variable, and 83 exemplars (matrix with $N = 83$ and $C = 10$). The database was collected from the Appendix of Gigerenzer and Goldstein (1996). The cues (feature values) in the simulation are at best approximate to the ones used by our participants. We have no way of knowing whether their inferences were actually based on these nine cue-dimensions. Rather, we take the matrix to be an estimate of the *similarity structure* that describes the environment of German cities. The assumption is: If two cities are similar to each other with regard to the nine cue-dimensions used in the simulation with PROBEX, on average they are similar also with respect to the cue-dimensions not coded in the simulation. Nevertheless, our imperfect knowledge of the knowledge-state of the participants provides a ceiling on the quantitative fit to behavioral data that can be expected from PROBEX.

We have more knowledge of what cities are known, because this was rated by the participants. If we score a "recognize response" with 1 and a "not-recognize response" with 0 we get an estimate of the retrieval probability p_n for city n. On average the participants reported to have recognized 29.4 ($SD = 12.9$), or 35% of the German cities. For example, Berlin was recognized by all 40 participants and thus gets $p_n = 1$, while Neuss was known to only one participant leading to a $p_n = 1/40$. These estimates of p_n were entered in the simulations with a separate retrieval probability for each city (i.e., in order to implement the recognition principle). For example, in each specific iteration of PROBEX, the probability that each cue value of Neuss

is as known 1/40. Unknown probes eliciting a default vector of "?" are thus similar to small known cities that also contain many "?-marks."

Because of the addition of a normally distributed response error, Eq. (8) may produce numbers outside the ranges of the probability scales. The output from Eq. (8) was truncated separately for the half-range and full-range probability scales. For the full-range format, outputs from Eq. (8) lower than 0 were assigned overt probability 0, and outputs larger than 1 were assigned overt probability 1. For the half-range format, outputs lower than .5 were assigned overt probability .5 and outputs above 1 were assigned overt probability 1.

PROBEX was applied to the knowledge-state matrix with four free parameters, the parameter s for similarity, the parameter ϕ for dampening, the parameter k for the stopping rule, and the parameter σ_r^2 for response error variance. These parameters were simultaneously fitted to the data from the half-range and full-range conditions. In the half-range condition, the dependent variables were the distribution of confidence judgments (6 data points, $df = 5$) and the proportions correct in each confidence category (6 data points, $df = 5$). In the full-range-condition, the dependent variables were the distribution of probability judgments (11 data points, $df = 10$) and the proportion of true statements in each probability category (11 data points, $df = 10$). In addition, the responses were used to determine the solution probability for each of the 40 items ($df = 40$) and they made 40 point estimates ($df = 40$). Altogether, this makes four free parameters fitted to 114 data points ($df = 110$).[7]

Best-fitting parameters were searched by an extensive and iterative grid-search procedure and RMSD as the error function. The model-fitting procedure is described in detail in Appendix A. The best-fitting parameter values were $s = .04$, $\phi = .9$, $k = .7$, and $\sigma_r^2 = .055$. The similarity parameter s implies a similarity-graded probability on the picky side. The estimate of the response error variance σ_r^2 is larger than previous estimates based on similar data (Juslin et al., 1997, 1999, 2000), perhaps a consequence of the unfamiliar topic of these tasks (to Swedish participants). The stopping parameter k implies that, on average, only about two exemplars are sampled before a judgment is made. These parameters (i.e., low s, high k) coincide with those that afford fast, accurate and robust performance in a state of limited knowledge.

Table 2 provides a summary of the data and the predictions by PROBEX for the dependent variables used in the model-fitting procedure, along with summary statistics for the fit in terms of RMSD and coefficients of determination r^2. Below, PROBEX, with the same parameters, is also used to predict calibration with interval estimation and response times.

5.2.2. *Prediction of point estimates*

In Fig. 5A, mean observed point estimates ($n = 40$) are plotted against predicted point estimates. The predictions by PROBEX account for 96% of the variance in the empirical point estimates (RMSD = 81,330). Fig. 5B and C plot the participants' and PROBEX' estimates, respectively, against the German city-populations. The correlation between the participants' median point estimates and true city populations was .95. The correlation between the predictions by PROBEX and city population was .97. Because both the participants' median estimates and PROBEX' predictions proved quite accurate, the residuals in these estimates are modest (110,463 and 85,200, respectively). However, there is a correlation between the errors in participants' and PROBEX' estimates ($r = .68$), illustrating that the participants and PROBEX share important misconceptions about the populations of the German cities. We

Table 2
Dependent measures for the simulations with PROBEX and the judgment data from the participants in the experiment

Dependent measure		Data set	
Judgment	Index	PROBEX	Participants
Point estimates ($n = 40$)	Mean (true, 304,976)	323,287	315,921
	Median (true, 179,562)	189,650	182,500
	Standard deviation	301,034	343,857
	Achievement (r)	.97 ($r^2 = .94$)	.95 ($r^2 = .90$)
	Mean square residual	85,200	110,463
	Fit of PROBEX	$r^2 = .96$, RMSD = 81,330 (correlation between residuals = .68)	
Solution probabilities ($n = 40$)	Mean	.68	.68
	Fit of PROBEX	$r^2 = .61$, RMSD = .11	
Probability judgments ($n = 34$)	Mean probability (HR)	.68	.68
	Proportion correct (HR)	.68	.68
	Over/underconfidence (HR)	0	0
	Mean probability (FR)	.48	.43
	Over/underconfidence (FR)	.09	.09
	Fit of PROBEX	$r^2 = .96$, RMSD = .06	

HR: half-range; FR: full-range, IE: interval estimation.

conclude that PROBEX provides a satisfactory fit to the point estimates and that both PROBEX and the participants make accurate point estimates, considering the limited knowledge and the retrieval of very few exemplars (approximately 2).

PROBEX allows a more detailed investigation of the processes. If we disable the recognition principle by making knowledge perfect for all recognized cities regardless of their size, the achievement is only .79, as compared to .97 when the recognition principle is implemented. The fit between PROBEX and the participants' point estimates is likewise lowered (from $r^2 = .96$ to .61). This suggests that the "recognition principle" is an important aspect of the processes that underlie the participants' judgments. Moreover, as might be expected, this reliance on recognition is less important for inferences that involve large cities. Removing the recognition principle for cities with large population has only a modest effect on the correlation between the predictions by PROBEX and the participants' judgments (i.e., decreasing from .98 to .92 when only the 12 largest cities with 250,000 or more inhabitants are entered into the calculation of the correlation). Evidently, large cities tend to be well known to the participants, leaving little room for the recognition principle to operate, and they have to rely on similarity-based inferences from known properties.

5.2.3. Prediction of solution probabilities

Fig. 5D plots empirical solution probabilities against predicted solution probabilities. Solution probability is defined as the proportion of participants that selected the correct answer to an item. In half-range, "correct" refers to a decision "True" for cities with a population above 180,000 inhabitants, and a decision "False" for cities with a population below 180,000

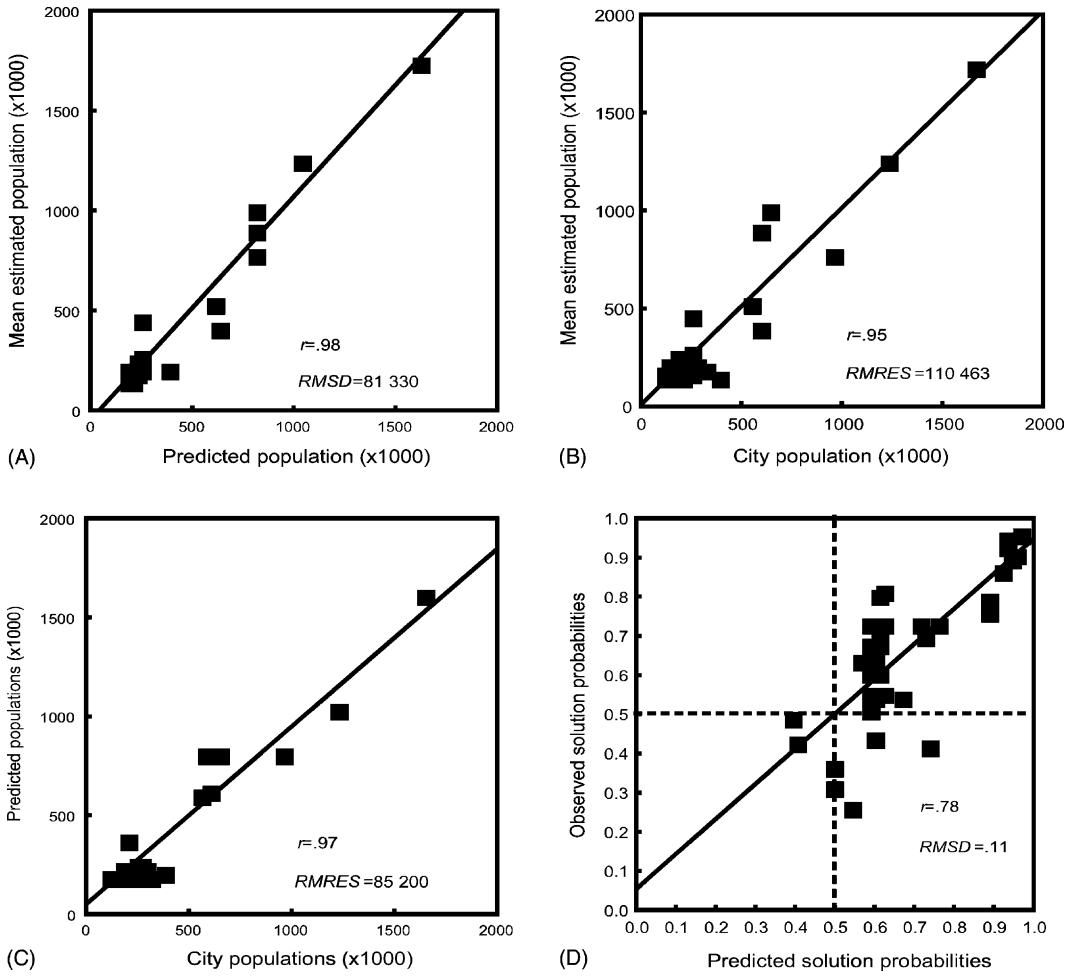


Fig. 5. (A) Mean point estimates of German city-populations by 40 participants plotted against the predictions by PROBEX. (B) Participants' mean estimates plotted against the true values (human achievement). (C) PROBEX' estimates plotted against the true values (PROBEX achievement). (D) The observed solution probabilities plotted against the predictions by PROBEX.

inhabitants. In the full-range condition, the solution probability is based on whether the probability response is in the proper end of the scale. If the statement is correct then probabilities above .5 are considered correct and if the statement is false then responses below .5 are correct. Probabilities of .5 were randomly scored as correct or wrong. The solution probabilities in Fig. 5D are means across both the half-range and the full-range condition (which are almost identical). Items with solution probabilities below .5 are cities for which most of the participants make an erroneous inference (so-called "misleading items"). Cities with solution probabilities above .5 are the cities for which most participants make a correct inference (cf. error-dependence, e.g., in Juslin & Olsson, 1997). Will PROBEX make the same errors as humans do?

Fig. 5D illustrates that there is a strong positive correlation ($r = .78$) between predicted and empirical solution probabilities, although, the fit is obviously poorer than for the point estimates (Fig. 5A). However, PROBEX correctly identifies 4 out of 7 misleading items and 33 out of 33 non-misleading items. (The correctly categorized items are those in the lower left and the upper right quadrants of Fig. 5D.) Specifically, both the participants and PROBEX misjudged the populations of Wuppertal, Bielefeld, Krefeld, and Erfurt. The explanation is that the recognition principle suggests the wrong answer for these cities. The correct classification of 37 out of 40 items is impressive given (a) the sampling error in the solution probabilities, and (b) our very limited knowledge of the specific cue-structures that underlie the judgments. The positive correlation between the observed and predicted solution probabilities nevertheless suggest that our database capture important aspects of the similarity structures used by the participants. Because PROBEX predicts of both confidence and solution probabilities—and thus what items that will be misleading—quite accurately, it will produce overconfidence bias when items are selected for difficulty or "misleadingness" (Gigerenzer et al., 1991; Juslin et al., 2000).We conclude that PROBEX is human-like not only in terms of its achievement, but also in regard to some of its misconceptions.

5.2.4. Calibration of subjective probabilities

A half-range calibration curve plots the proportion of correct decisions as a function of the subjective probability assigned to the decision. To the extent that the confidence judgments are calibrated (realistic), the proportions fall on the identity line ("ideal" calibration in Fig. 6A). Fig. 6A presents observed and predicted calibration curves for the half-range condition. Fig. 6C provides the distributions across the six confidence categories (i.e., collapsed across items and participants).

Observed mean confidence in the half-range condition was .68 (95% confidence interval, CI $= \pm.01$), which coincides with the overall proportion of correct answers, .68 (95% CI $= \pm.02$). There was thus no general bias and the over/underconfidence score was 0 (95% CI $= \pm.02$). The miscalibration in Fig. 6A mainly reflects the regression arising from stochastic components of the judgment process (see Erev et al., 1994), such as response error in the use of the probability scale (Juslin et al., 1997, 1999). Specifically, the regression is caused by the end-effects imposed by the probability scale. "True" degrees of belief equal to 1 can only lead to overt probabilities equal to or lower than 1 after the addition of a response error, while we have the reverse effect at the other end of the scale. The predictions by PROBEX reproduce the same level of over/underconfidence (0) as in data, with identical mean confidence (.68) and proportion correct (.68). Despite some idiosyncrasies, the predicted calibration curves capture the overall shape of its empirical counterpart.

Fig. 6B presents data and predictions for the full-range condition where the participants assessed the probability that a statement is true on a scale between 0 and 1.0. In Fig. 6B, the *x*-axis is therefore scaled from 0 to 1. A second difference from Fig. 6A is that the proportions on the *y*-axis in Fig. 6B are not specifically proportions *correct* (i.e., no choice precedes the full-range assessment), but the proportion of the statements that are true. As in Fig. 6A, however, perfect calibration is a calibration curve that falls on the identity line. Fig. 6D presents the distribution over the 11 probability categories.

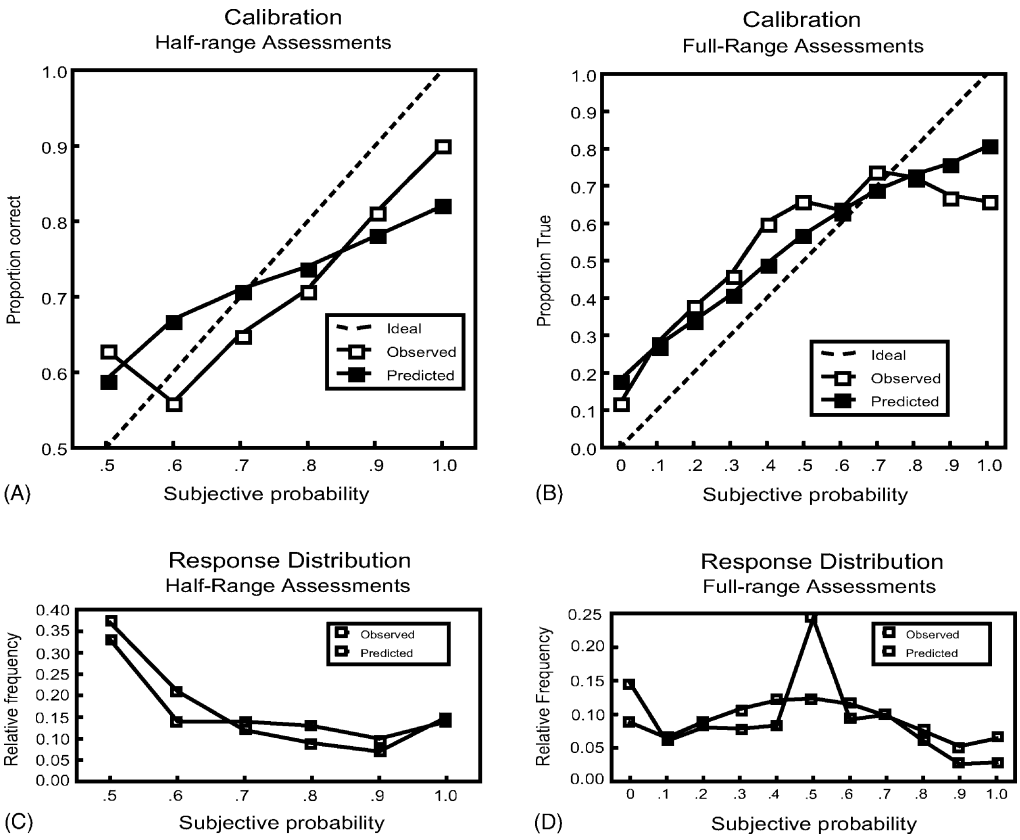


Fig. 6. (A) Observed calibration curve for the half-range condition and predictions by PROBEX. (B) Observed calibration curve for the full-range condition and predictions by PROBEX. (C) Observed distribution across the six categories of the half-range probability scale and predictions by PROBEX. (D) Observed distribution across the 11 categories of the full-range probability scale and predictions by PROBEX.

The observed calibration curve in Fig. 6B that suggests modest calibration has two salient features. First, there is an *underestimation* of the probability that the city has a population that exceeds 180,000 inhabitants (i.e., this is reflected in a calibration curve above the diagonal).[8] By experimental design 50% of the statements were true (180,000 is the median city population), but the mean probability was .43 (95% CI $= \pm .01$). Second, the curve drops for the highest subjective probabilities in a rather perplexing way.

The predictions by PROBEX are presented in Fig. 6B and D, and in Table 2. PROBEX parallels the under-estimation of the probability that a city has more than 180,000 inhabitants, although to a lower extent than in humans. The predicted mean probability is .48. The explanation is that many probes are unknown to the participants (and to PROBEX). Because of the correlation between recognition and population, the populations of vaguely known exemplars are likely to be small. PROBEX therefore assigns new unknown cities a low probability of having a large population. The reason why this tendency is stronger in the participant data is presumably that the participants draw on a larger and even more biased exemplar knowledge

base than PROBEX including other, primarily very large European cities, like London and Paris. The slope of the calibration curve, again disclosing regression due to the response error, is similar for the data and PROBEX.

On the other hand, there are two deviations between predictions and data. First, although the calibration curve produced by PROBEX for the full-range condition flattens out somewhat in the highest probability categories, PROBEX fails to reproduce the drop in the empirical calibration curve. At present, we have no explanation for this drop in the calibration curve, but we propose that it reflect some idiosyncrasy in the cue-structures that underlie the performance of the participants. Second: In the observed response distribution there is a peak in the .5 category that is not captured by the predictions. One interpretation is that the .5 responses are a mix of two different responses, inferential responses (modeled by PROBEX) and truly random decisions made without any attempt to make an inference. ("I know nothing at all about German cites, so I respond 50/50.") A similar distinction is explicitly modeled in the *combined error model* (Juslin et al., 1997, 1999).

In the interval estimation condition, finally, the participants assess intervals intended to include the true value with probability 1 (100% certainty). As often observed in previous studies, there was gross overconfidence with the interval estimation format. The 1.0 intervals only included the true values with an observed proportion of .72 (95% CI $= \pm .05$).

The over/underconfidence with interval estimation expected from the full-range calibration curve generated by PROBEX was approximated in the following way. The expected proportion falling within the 1.0 probability interval was estimated by computing the difference between the proportion of true statements in subjective probability category 1.0 minus the proportion of true statements in subjective probability category 0. This is the expected proportion within the 1.0 probability interval if the participants disclose the same calibration in their placement of the upper (1.0 fractile) and the lower (0 fractile) limits as they do in their full-range assessment. The predicted proportion falling within the 1.0 interval is 62, implying a an overconfidence of .38 (i.e., $1 - .62$). The corresponding observed proportion in the interval estimation condition was .72 implying an overconfidence of 28.

One account for this difference is that the predictions derived from the full-range calibration curve are more affected by the scale end-effects imposed by the [0, 1] limit of the probability assessment scale that contribute to the observed overconfidence (see Note 3). With interval estimation, assessments are made on the population dimension that has less salient scale limits and the scale-end effects do not contribute to overconfidence in the same systematic manner. At present, however, this explanation remains a speculation.

Observed over/underconfidence scores are presented in Fig. 7 with 95% confidence intervals. Fig. 7 reveals the enormous format-dependence in data. When the participants are probed for their confidence with the half-range format over/underconfidence is zero, but when the same content is approached with interval estimation the overconfidence is enormous (.28). The format-dependence effect is qualitatively reproduced by PROBEX and explained by the response error included in the response stage of the model (see Juslin et al., 1999).

5.2.5. Response times

Recall that when PROBEX is applied to the present task, it was assumed that the basic stopping rule concerns a point estimate (Eq. (9)). This captures the intuition that the prime

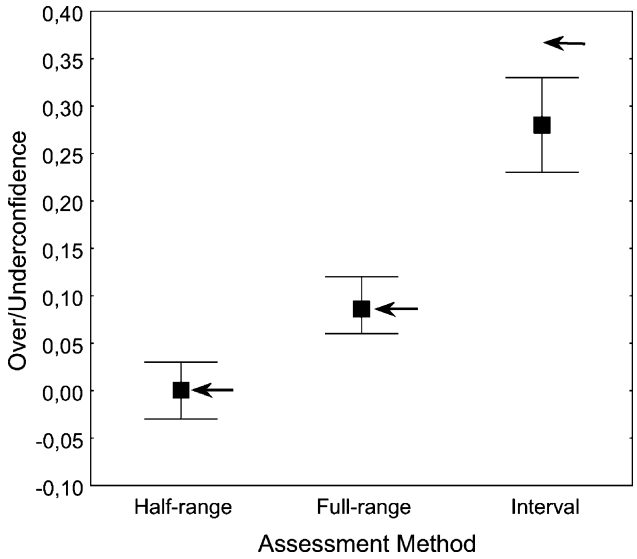


Fig. 7. The format-dependence effect observed in the data, where the over/underconfidence score is plotted for each assessment format. The arrows indicate the corresponding predictions by PROBEX.

concern of the participants in this task is to get a clear conception of the population of each presented German city. The analysis of the empirical response times revealed that there was little systematic variation in the response times for different cities. It turned out that with parameters fitted to the point estimates, decisions, and probability judgments, PROBEX reproduced this pattern. PROBEX (and presumably the participants) have little knowledge and thus uniformly retrieve a small number of exemplars.

This uniform pattern is illustrated in Fig. 8 that presents median observed and median predicted response times as a function of subjective probability. Predicted response times in Fig. 8A are re-scaled by multiplication with $m(\mathrm{RT})/m(N)$, where $m(\mathrm{RT})$ is the mean across the six observed data points in Fig. 8A and $m(N)$ is the corresponding six predictions (i.e., the mean number of sampled exemplars). The same procedure was applied to the 11 values in Fig. 8B. This serves to re-scale the predictions into milliseconds.

When mean response time is plotted against probability category, both the participants and PROBEX (with the current stopping rule) produce essentially flat functions (Fig. 8). This is expected if the stopping rule is formulated in terms of the point estimate which bears no obvious relation to subjective probability. One systematic discrepancy, however, is that for half-range probability judgments, there is a tendency for the mean response times to fall with increasing subjective probability. It appears that with this format the participants have turned to a stopping rule defined in terms of the two-alternative decision made prior to the half-range probability judgment rather than the point estimate. A falling function is typical of stopping rules that are formulated in terms of a binary decision (i.e., as presumed by most sequential sampling models, e.g., Juslin & Olsson, 1997). This tendency is weak, though, again presumably because the participants uniformly retrieve few exemplars. Although PROBEX accounts for some aspects of the data, future research should more

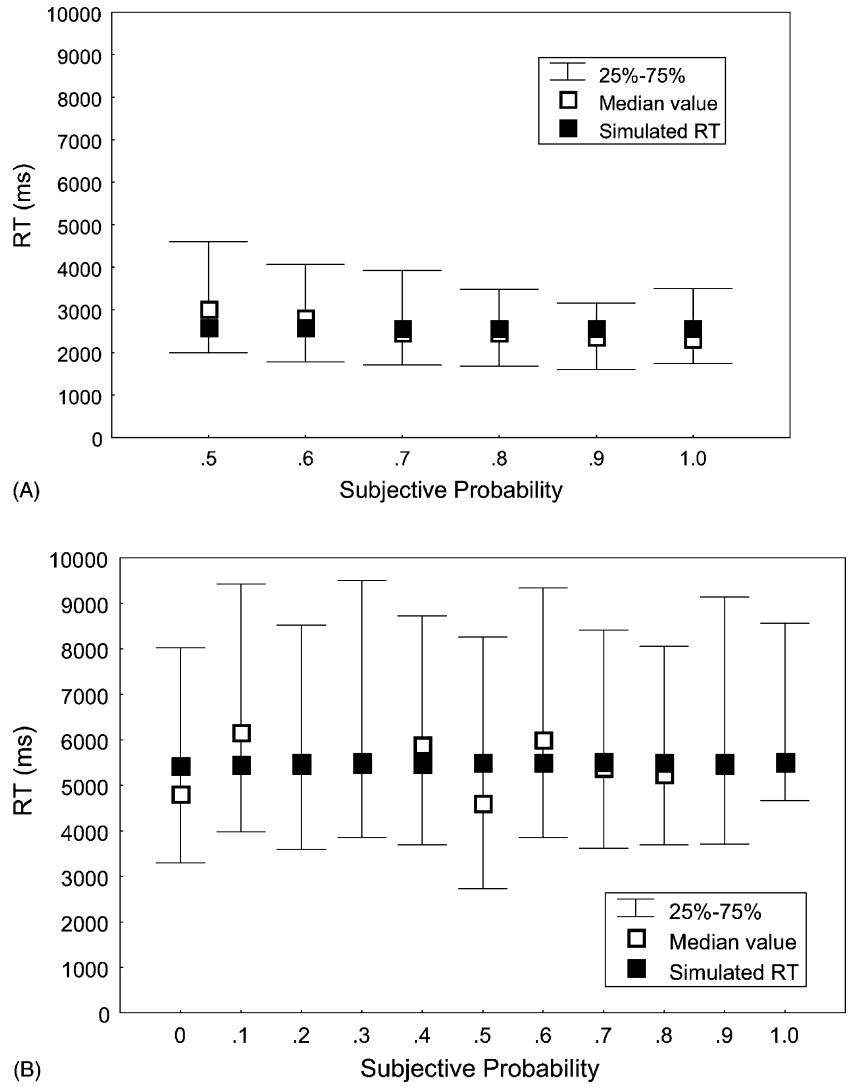


Fig. 8. (A) Observed and predicted response times as a function of subjective probability category in the half-range condition. (B) Observed and predicted response times as a function of subjective probability category in the full-range condition. Because the distributions of response times are extremely skewed they are presented in terms of medians and the .25 and the .75 percentiles.

carefully ascertain the stopping rules that best account for response times in generic knowledge tasks.

5.3. Discussion

Overall, we conclude that PROBEX reproduces the basic patterns in data. Point estimates by PROBEX coincide quite well with those by human participants, reproducing the same overall

achievement but also many of the misconceptions. At the same time, PROBEX reproduces empirical calibration curves and probability distributions with decent fidelity. Because PROBEX considers the effects of a random response error in the response process (Juslin et al., 1997, 1999) it mimics the simultaneous over- and underconfidence bias with different probability assessment formats. Finally, with parameters fitted to point estimates, decisions and probability judgments (but *not* to response time data), PROBEX reproduces also surprising aspects of the response times (i.e., the flat functions in Fig. 8).

The quantitative fit is impressive considering that: (a) PROBEX simulates the entire process whereby generic knowledge is transformed into observable judgments. (b) The model *simultaneously* predicts point estimates, binary decisions, overt subjective probability assessments with three different assessment formats, and response times. While a model may achieve a fortuitous fit to a single dependent variable (Roberts & Pashler, 2000), the simultaneous consideration of multiple dependent variables puts non-trivial constraints on any aspiring cognitive mechanism. (c) Yet, the simulations in this article are based on a very crude model of the database and the specific cue-structure used by the participants. It is significant that the main departures between predictions and data occur for the solution probabilities (Fig. 5D), the predictions that most clearly depend on knowledge of the specific cues that support the judgments. (d) Four free parameters were fitted to 114 data points.

Moreover, PROBEX allows us to explain a number of more specific aspects of the process. For example, why Swedes underestimate the probability that German cities have more than 180,000 inhabitants and why the response times are flat functions of probability category. (Both are mainly explained by limited knowledge and reliance on the recognition principle.) Although these peculiarities may appear uninteresting *per se*, the fact that PROBEX map these and other idiosyncrasies that surface when Swedes approach the German city-population task suggests that it makes serious contact with the processes that compute these judgments. The poor fit of PROBEX when the recognition principle is disabled ($r^2 = .79$ vs. .96) further testifies to the importance of the recognition principle.

The best-fitting parameters indicate a fast-and-frugal version of PROBEX: the participants infer properties of new objects by rapid retrieval of a small number (approximately 2) of highly similar exemplars. PROBEX and the participants nevertheless make good estimates. In sum: this section establishes that an exemplar model like PROBEX is a viable candidate theory of the processes that make probabilistic inferences from generic knowledge. Quantitative fit is, of course, a weak way of establishing the validity of a model and we need further strong tests of parameter-free predictions, comparisons with other models, and applications to more complex real-life like environments. Yet, tentatively it is both striking and encouraging that the best-fitting parameters suggests a fast-and-frugal version.

6. General discussion

6.1. Summary

A plausible model of the cognitive processes that make inferences and compute “degrees of belief” should have, at least, three properties. *First*, it should be consistent with, and preferably

extend on, models with independent support in the *Cognitive Science* literature. PROBEX inherits its fundamental psychological principles from one of the most successful models—the context model (Medin & Schaffer, 1978; Nosofsky, 1984). In the first section, we pointed out that, if the output of the context model is interpreted as subjective probabilities it predicts that subjective probabilities should be responsive solely to frequencies in some environments and solely to similarity in others. Moreover, the model embodies a "hybrid-representation," similarity-graded probability, which provides a useful compromise between demands for relevance and sample size, especially when knowledge is scarce.

In the second section, the context model was developed into an algorithm that computes subjective probabilities from generic knowledge. PROBEX relies on similarity relations between retrieved rather than visually presented features, adds a sequential sampling mechanism, and compute point estimates, decisions, subjective probabilities, and response times. By dispensing both with assumptions of obligatory and complete storage, and exhaustive retrieval of all exemplars, PROBEX attempts to implement an exemplar algorithm that better satisfies the desiderata of bounded rationality (Simon, 1990). The inferences can thus be made by speeded retrieval of a small number of similar exemplars.

The second property of a plausible model is addressed by research on ecological rationality (Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1999). The algorithm should exploit structures of real environments, while making psychologically plausible demands on time, knowledge, and computational resources. The discussion of frugality in Gigerenzer et al. (1999) was complemented by considerations of the demands for pre-computed knowledge. In contrast to TTB and multiple linear regression, PROBEX is a lazy algorithm requiring no pre-computed knowledge. The frugality of PROBEX is controlled by the number of exemplars sampled to make a judgment. Simulations demonstrated that, when PROBEX relies on all exemplars—as TTB and multiple regression always does, it dominates over its competitors, in particular when knowledge is scarce. When PROBEX retrieves a few very similar exemplars, it is particularly efficient, thereby implementing a fast-and-frugal exemplar model.

The third important aspect of a cognitive model is that it validly describes the processes that compute "degrees of belief" from generic knowledge. In Juslin et al. (2001), PROBEX proved to be an accurate model of how people respond to similarity and frequency in a category learning task. In the final section of this article, PROBEX was applied to the point estimates, decisions, subjective probabilities, and response times generated by participants in a generic knowledge task. The fit was satisfactory and the best-fitting parameters suggested that the participants adapt to the demand for probabilistic inference from generic knowledge by retrieval of a few similar exemplars. Responding with such a picky, similarity-graded probability is especially efficient in states of limited knowledge.

Because a picky, similarity-graded probability comes close to a pure relative frequency computation, this suggests some resemblance to models that emphasize relative frequency (e.g., Björkman, 1994; Gigerenzer et al., 1991; Juslin, 1994; Soll, 1996). In contrast to these models, however, PROBEX does not pre-compute relative frequencies conditional on single cues (e.g., the relative frequency of German cities with more than 180,000 inhabitants in the subset of cities that have a university). PROBEX retrieves exemplars with large similarity to the probe, computing a picky similarity-graded probability on the spot.

These results reported in this article accumulate to the conclusion that PROBEX provides a viable candidate theory of the processes that compute probabilistic inferences from generic knowledge. In the remaining part of this section, we comment on the theoretical key-properties of PROBEX and discuss its relations to extant models.

6.2. Fast and frugal exemplar models

One criticism of exemplar models is that they make too excessive storage demands (e.g., Barsalou et al., 1998; Nosofsky et al., 1994), in particular, models that presume obligatory storage of each exposure to an object. Although such claims are notoriously difficult to evaluate without the help of some comprehensive theory of the costs involved in storage and computation, we note that PROBEX is an exemplar model that makes no assumption of obligatory storage. Moreover, the extent of exemplar retrieval is an empirical issue in terms of the parameter *k*. Yet, PROBEX is an efficient tool for inference.

This suggests an important research program: to explore exemplar models that make demands that are more modest on storage and retrieval. For example, can exemplar models that presume storage of only a subset of objects, or retrieval of only a few exemplars, provide as good fit to classification data as more demanding versions? What are the boundary conditions? Successful application of fast-and-frugal exemplar models would increase their plausibility by bringing them within the scope of bounded rationality (Simon, 1990).

6.3. Lazy algorithms

On several occasions in this article, we have stressed that one virtue of PROBEX is that it is a lazy algorithm (Aha, 1997): it requires no pre-computed representations, like beta-weights or cue validities. One way to capture this virtue is by saying that PROBEX obeys a *principle of minimal commitment*. The best way to minimize storage demands and maximize flexibility is by not committing to premature and complex computation of abstractions, but to postpone computation to the time when the task is known. In an environment with unforeseen task demands, this minimizes the need to compute and store abstractions.

Interestingly, this argument reverses the argument of "cognitive economy" routinely advanced in favor of abstractions, such as rules and prototypes (e.g., Rosch, 1978; Smith & Minda, 2000), that the storage of a simple rule or prototype obviates the need to store single instances of a category. The contradiction is only apparent, however. When future task demands are known, computation of abstractions minimizes, if not the computational demands, at least the storage demands. When knowledge is for flexible and opportunistic use in novel situations, the principle of minimal commitment is more appropriate. Arguably, in many real-life contexts, the latter kind of "economy" is the more relevant one.

One solution to these contradictory demands is to rely on several knowledge systems that each work along different principles, for example, one abstract, rule-based system and one exemplar-based system. The accumulation of data that requires models incorporating multiple levels of representation (e.g., Ashby, Alfonso-Reese, Turken, & Waldon, 1998; Erickson & Kruschke, 1998; Logan, 1988) suggests that this may indeed be the solution favored by evolution.

6.4. A mechanism for the recognition principle

The recognition principle states that if one but not the other city is recognized, one should guess that the recognized city has a larger population. One consequence of this principle is that accuracy may actually decrease with increasing knowledge (viz., "The less is more effect" in Gigerenzer & Goldstein, 1996). In the simulations by Gigerenzer and Goldstein, accuracy was higher when 50–75% of the cities were known than when most or all cities were known. If all cities are known, the algorithms cannot benefit from the powerful recognition principle. In a binary choice task, the principle can be modeled as a binary cue, but extension to other judgment tasks requires a more general formulation.

PROBEX presumes no explicit check of a recognition cue (Gigerenzer & Goldstein, 1996): the principle *emerges* as an integral aspect of the exemplar algorithm. Intuitively, this corresponds to the experience that you recognize a city but little is known about it. Hence, it is similar to other cities you know little about and these tend to have a small population. This furnishes a cognitive mechanism and explicates the intuition that recognition is a "probability cue" among others. Moreover, this mechanism implies that the principle is flexible, operating only when recognition is related to the criterion, as for city populations. If the vaguely known exemplars tend to have *large* values, this mechanism elicits an inference that goes contrary to the original recognition principle. That is, unrecognized objects will be expected to have large criterion values. This suggests that the recognition principle need not be an explicit part of the decision process, but a flexible and emergent property of exemplar architectures.

6.5. Relations to other models

PROBEX is obviously related to several models of perceptual classification, including, the context model (Medin & Schaffer, 1978; Nosofsky, 1984) and EBRW (Nosofsky & Palmeri, 1997). In the exposition of PROBEX, we have repeatedly pointed out the heritage from these and other, related models. However, none of these models applies to probabilistic inference from generic knowledge or probability judgment and, therefore, they are not competing accounts of the data presented in this article. Indeed, we have been unable to find extant models with the same broad coverage of dependent variables that allow a similar in depth analysis of the processing in a generic knowledge task. Nonetheless, a number of models address issues that partially overlap with those of concern to PROBEX.

6.5.1. TTB

TTB can potentially account for the binary decisions and half-range probability judgments on the assumption that the participants report the cue validities as their subjective probability judgments (cf. Gigerenzer et al., 1991). Although, TTB originally applies to pair-comparisons, the refinements needed to apply TTB to the data from the single-object tasks in the experiment reported above should be manageable and straightforward.

However, TTB does not predict point-estimates, probability judgments with other formats, or response times. Except for point estimates, this is true even if the scope is extended to the entire "*adaptive tool box*" of heuristics envisioned by Gigerenzer et al. (1999). Moreover, the idea of an adaptive tool box containing a large number of disparate heuristics, each appropriate

to a specific judgment format (e.g., TTB for pair comparisons, QUICKEST for point estimates), suggests a fragmented view of the belief system. In contrast, PROBEX captures the intuition that different judgment formats, like the ones in the experiment reported above, all revolve around a common core of beliefs about the environment.

Although people sometimes rely on lexicographic decision rules in multi-attribute decision tasks where the information is presented on an information board (Payne, Bettman, & Johnson, 1990; Tversky, 1969) and people have been observed to rely on TTB in specific circumstances (Bröder, 2000), there are problematic aspects of data. A recent study (Jones, Juslin, Winman, & Olsson, 2000) provided little support for TTB in a standard multiple-cue judgment tasks. TTB cannot handle non-linearly separable categories (the *EXOR*-problem), but humans can (Kruschke, 1992). This capacity requires consideration of more than a single cue, and is thus troublesome for all algorithms of the TTB-family. Finally, it is hard to see how TTB accommodate several phenomena investigated with exemplar models, like the power-law speedup in expertise (Logan, 1988; Nosofsky & Palmeri, 1997; Palmeri, 1997).

One possibility is to regard TTB and PROBEX as complementary, in the spirit of models with multiple representation levels (Ashby et al., 1998; Erickson & Kruschke, 1998). Perhaps, TTB is fast and frugal within the realm of a rule-based system constrained by working memory capacity, while exemplar models employ fast and frugal retrieval processes adhering to the principle of minimal commitment. This eclectic attitude non-withstanding, at present there is little data demonstrating that people frequently rely on TTB.

6.5.2. *Combined error model*

The combined error model accounts for a number of phenomena in the probability judgment literature, including the overconfidence phenomenon, the hard–easy effect, conservatism, base-rate effects, expertise effects, and format dependence (Juslin & Olsson, 1999; Juslin et al., 1997, 1999). The model extends on the ecological models of confidence (Gigerenzer et al., 1991; Juslin, 1994) by modeling the stochastic components of the judgment process (Erev et al., 1994). The model fits calibration curves and probability distributions, but not point estimates and response times.

As mentioned above, the picky frequentist version of the context model (Eq. (3)) computes the internal probabilities of the combined error model. This means that in essence the combined error model is a special case of PROBEX (i.e., with $k = 0$, $s = 0$, and $\phi = 0$, implying exhaustive sampling, a picky frequentist, and no dampening). The difference is that, in the combined error model, sampling error in experience and retrieval is modeled explicitly by a binomial distribution, whereas this error is implicit in the application of PROBEX to the incomplete knowledge matrix. This difference reflects that PROBEX is an explicit account of how the subjective probabilities are computed from generic knowledge.

The sampling and response errors that provide the combined error model with its explanatory power are thus integral aspects also of PROBEX. The fact that PROBEX inherits crucial properties from the combined error model implies that it predicts additional phenomena not discussed in this article, like conservatism and hard–easy effects. On the other hand, PROBEX widens the scope of the combined error model by allowing degrees of belief to be affected by similarity, and by providing point estimates and response times. Thus, the combined error model should be able to account for the calibration curves and probability distributions to a

degree that approximates the fit of PROBEX. However, it cannot account for the point estimates and the response times, or for the role of similarity.

6.5.3. MINERVA-DM

A recent model that is similar to PROBEX is MINERVA-DM (Dougherty et al., 1999). Both models apply exemplar (or instance) models to judgment. MINERVA-DM can fit decisions and subjective probability judgments and should be able to account also for point estimates with minor modifications. MINERVA-DM does not predict format dependence or response times, but could be amended in suitable ways.

Although PROBEX and MINERVA-DM are similar in some respects, there are important points of difference. First, in contrast to MINERVA-DM, PROBEX embodies a sequential sampling of exemplars. The sequential sampling mechanism allows PROBEX to implement a fast-and-frugal exemplar model that computes accurate judgments while making psychologically plausible demands on time, knowledge and computation (Gigerenzer et al., 1999). It also means that PROBEX can fit response time distributions. Although a minor aspect of the present article, the use of multiple dependent measures, including response times, may prove crucial in the triangulation to identify the cognitive processes.

Second, whereas the exemplars in PROBEX refer to "crystallized" knowledge structures corresponding to objects in the environment, MINERVA-DM presumes that each encounter with an object leaves a unique memory trace. The exemplars in PROBEX do not rule out that the knowledge structures themselves have been formed by a variety of cognitive processes, including processes that involve abstraction across stimuli. Applying MINERVA-DM to a generic knowledge task like the German city-population task is less than straightforward. For example, one will have to make assumptions about the frequency of German cities to generate predictions. This difference is theoretically very interesting and well worthy of investigation in future research, but we note that there exists preliminary evidence in favor of the representational scheme used by PROBEX (Barsalou et al., 1998).

The third difference is one of emphasis. MINERVA-DM is primarily applied to cognitive biases in the judgment literature, routinely explained by similarity computations (i.e., the representativeness heuristic). MINERVA-DM is applied to abstract and arbitrary sequences of vectors with little relationship to real environments. This application accords with the content-blind and abstract application in the memory and categorization literature. PROBEX, on the other hand, takes ecological rationality as point of departure. Similarity-graded judgment is therefore primarily viewed as an efficient tool for inference.

Finally, we depart from some of the claims by Dougherty et al. (1999) concerning MINERVA-DM. We do not agree that exemplar models like MINERVA-DM or PROBEX predicts base-rate neglect in any natural way. The base-rate neglect predicted by MINERVA-DM arises from the auxiliary assumption that people confuse posterior probabilities and likelihoods in Bayesian problems, not from MINERVA *per se*. Adding this assumption to many other models should likewise lead to prediction of base-rate neglect.

Likewise, although exemplar models respond to similarity, the conjunction fallacy is not a natural consequence of an exemplar model. For example, inspection of Eq. (2) of the original context model makes it evident that straightforward application of the model predicts judgments that conform with the conjunction rule. Only if the extension of concepts (e.g., feminists,

or bank-tellers) is misrepresented will an exemplar model like MINERVA-DM predict the conjunction fallacy (e.g., by assuming that people presented with the probe "bank teller" retrieve only bank tellers that are not feminists). Again, misrepresented extensions can be assumed by a multitude of other models with similar results (and one may sense a threat of circularity with the entire approach). In sum: several of the phenomena discussed by Dougherty et al. (1999) are not naturally implied by exemplar models, like the context model or *MINERVA-2*, but arise from auxiliary assumptions.

6.6. Conclusions

In this article, we have proposed that the context model provides a useful way to elucidate the role frequency and similarity in judgment. We have suggested that once that we take the pre-computation assumed by different algorithms into account, a "lazy algorithm" like an exemplar model may prove to be a more plausible candidate than acknowledged in previous research (e.g., Gigerenzer et al., 1999). Although we note that exemplar models are well supported by previous research in *Cognitive Science*, we stress that to fulfill the vision of a boundedly rational exemplar model, the validity of a specific algorithm like PROBEX needs to be validated by research that involves parameter-free predictions, comparisons between models, and application to more complex real-life like environments.

Notes

1. The original context model for binary features (Medin & Schaffer, 1978) is a special case of the generalized context model developed by Nosofsky (1984, 1986) that also handles continuous feature dimensions. The algorithm of PROBEX is actually based on the generalized context model and is thus equally apt at processing continuous feature dimensions. In the present article, however, we restrict ourselves to binary features.
2. This also means that PROBEX will disclose "The less is more effect" discussed by Gigerenzer and Goldstein (1996). The effect is that as the system gains more knowledge of the environment, implying a decreasing possibility to rely on the powerful recognition principle (or cue), the accuracy of the inferences may actually decrease.
3. The point estimate $c'(\bar{t})$ for continuous quantities could also be amended with a similar kind of dampening representing the equivalent of a "prior distribution" for the value of the variable, and with a similar random response error. We have chosen not to add this complication to the point estimates for continuous variables for two reasons. The assessment of the subjective probability $P(\bar{t} \in \mathrm{A})$ is especially sensitive to the problem of small sample sizes, and estimates without dampening become truly pathological at very small sample sizes (e.g., always 1 or 0 at sample size 1). Second, because the probability scale has salient end-points, the introduction of a random response error—even if it is has zero expectation—leads to quite pervasive scale-end effects (see, e.g., Juslin et al., 2000). Thus, whereas it seems crucial to model these aspects in the case of probability assessments, these components have less profound implications for point estimates for continuous quantities. Nevertheless, such amendments could be considered

also for the point estimates for continuous quantities in circumstances where it seems appropriate and useful.

4. Ridge regression has the drawback of biasing the predictions towards the mean, and thus lowers the predictive accuracy when the weights are calculated from many observations without problems with correlated variables. We hand-picked an intermediate ridge constant that increased accuracy with limited information (small training set) but did not lower performance with much information (large training set).
5. Note that these results for the overall proportion correct are dependent on the specific sampling of the dimension, training set size. In the simulations reported here, the smaller training set sizes are more densely sampled than the larger training set sizes (for the larger training set sizes, the function for proportion correct is virtually flat for most algorithms, see Fig. 1A). Another sampling of the training set-size dimension, for example, a uniform sampling, will thus produce a different proportion correct. However, because PROBEX yields the highest proportion correct regardless of the training set size (see Fig. 1A), it will always have the highest proportion correct.
6. We did not implement QUICKEST exactly as Hertwig et al. (1999) did as we did not use approximations to natural numbers which probably imply that our implementation gives slightly better predictions.
7. The 114 data-points are not strictly independent, even if we control for the degrees of freedom for each dependent variable. Specifically, two of the dependent variables, solution probability and calibration, are related by their algebraic dependency on the binary decisions elicited from the participants (i.e., on whether the decision was correct or not). While this complicates a statistical analysis, it has little implication for the interpretations made here. The ratio of the number of observations to the number of fitted parameters is still satisfactory.
8. The *overestimation* of the probability that the presented statement is true should be distinguished from the *overconfidence bias* that may be observed with half-range assessments. The former bias represents an overall tendency to believe that the presented statements are true, where as the overconfidence bias refers to an overestimation of one's own ability to select the correct answer.

Acknowledgments

This research was supported by the Swedish Council for Research in the Humanities and Social Sciences. The authors are indebted to the late Mats Björkman, Nick Chater, Michael Dougherty, Klaus Fiedler, Gustav Gredebäck, Sari Jones, Henrik Olsson, Pia Wennerholm, and Anders Winman for valuable comments on earlier versions of this article, and to Magdalena Jansson for running the experiment reported in the article.

Appendix A. The procedure used to fit PROBEX to data

The four parameters were varied in a grid search. The total number of parameter sets was $N = N_s \times N_k \times N_\phi \times N_{\sigma^2}$. One simulation was made for each parameter set, and the *RMSD* was

calculated between the simulated results and the empirical data for all eight dependent variables (four each in the half-range and full-range conditions, respectively). A matrix X_{ij} was filled with these RMSD-values, where each row (index i) corresponded to the eight RMSD-values from one parameter set, and each column (index j) was all N RMSD-values for one dependent variable. In order to decide which parameter set that gave the best fit to all dependent variables it was necessary to somehow normalize each column to make them comparable. It is not possible to sum the RMSD in a row and pick the row/parameter set with the lowest RMSD, because the city estimates would overshadow all the other since they are in the magnitude of hundreds of thousands. Thus, each column/variable was normalized by the smallest value in that column. This makes some values in the matrix 1, but most of them slightly greater.

$$Y_i = \frac{\sum_{i=1}^{8}(X_{ij}/\min(X_j))}{8}, \quad i \in [1, N] \tag{A.1}$$

Eq. (A.1) gives a measure Y_i for each parameter set and the parameter set with the smallest measure was selected as the best-fitting parameter set.

In order to illustrate that this method is reasonable consider the special case where the minimum found for each variable is in the same row. The normalizing procedure would then be unnecessary because it is obvious that this parameter-set provides the best fit. If this row is normalized and divided by 8, as prescribed in Eq. (A.1), the measure Y_i would be 1, the lowest possible value of Y_i. This demonstrates that the measure Y_i indicates of how well each parameter set minimizes the error for each dependent variable.

Because the simulations of PROBEX have random components, one grid search cannot find the "best" minimum. Therefore, we repeated our simulations both to narrow the range of parameters tested and to find stable parameter values.

References

Aha, D. W. (1997). Editorial. In D. W. Aha (Ed.), *Lazy learning*. Dordrecht: Kluwer Academic Publishers.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Andersson, J. R., & Fincham, J. (1996). Categorization and sensitivity to correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 257–259.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154–179.

Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442–481.

Barsalou, L. W., Huttenlocher, J., & Lamberts, K. (1998). Basing categorization on individuals and events. *Cognitive Psychology*, *36*, 203–272.

Björkman, M. (1994). Internal cue theory: Calibration and resolution of confidence in general knowledge. *Organizational Behavior and Human Decision Processes*, *58*, 386–405.

Brehmer, B. (1994). The psychology of linear judgment models. *Acta Psychologica*, *87*, 137–154.

Bröder, A. (2000). Assessing the empirical validity of the Take-The-Best-heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1332–1346.

Czerlinsky, J., Gigerenzer, G., & Goldstein D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 97–118). New York: Oxford University Press.

Dawes, R. M. (1982). The robust beauty of improper linear models in decision making. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 391–407). New York: Cambridge University Press.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968–986.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–528.

Estes, W. K. (1994). *Classification and cognition*. Oxford: Clarendon Press.

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*, 107–140.

Garg, R. (1984). Ridge regression in the presence of multicollinearity. *Psychological Reports*, *54*, 559–566.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic 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.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation: Letting the environment do the work. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hintzman, D. L. (1984). MINERVA-2: A simulation model of human memory. *Behavior Research Methods, Instruments and Computers*, *16*, 96–101.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple trace memory model. *Psychological Review*, *95*, 528–551.

Jones, S., Juslin, P., Olsson, H., & Winman, A. (2000). Algorithms, heuristics or exemplars: Process and representation in multiple-cue judgment. In L. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

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

Juslin, P., & Olsson, H. (1997). Thurstonian- and Brunswikian-origins of uncertainty in judgment: A sensory sampling model of confidence in sensory discrimination. *Psychological Review*, *104*, 344–366.

Juslin, P., & Olsson, H. (1999). Computational models of subjective probability calibration. In P. Juslin & H. Montgomery (Eds.), *Judgment and decision making: Neo-Brunswikian and process tracing approaches*. New York: Erlbaum.

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian- and Thurstonian-origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189–209.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format-dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1038–1052.

Juslin, P., Winman, A., & Olsson, H. (2000). Naive empiricism and dogmatism in confidence research: A critical examination of the hard–easy effect. *Psychological Review*, *107*, 384–396.

Juslin, P., Nilsson, H., & Olsson, H. (2001). Where do probability judgments come from? Evidence for similarity-graded probability. In J. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*, 136–153.

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

Klayman, J., Soll, J., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence? It depends on how, what and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*, 216–247.

Klein, G. A. (1989). Recognition primed decisions. *Advances in Man–Machine Systems Research*, *5*, 47–92.
Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44.
Lamberts, K. (2000). Information–accumulation theory of speeded categorization. *Psychological Review*, *107*, 227–260.
Logan, D. G. (1988). Towards an instance theory of automatization. *Psychological Review*, *95*, 492–527.
Maddox, W. T., & Ashby, G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, *53*, 49–70.
McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 128–148.
Medin, D. L., & Schaffer, M. M. (1978). Context model of classification learning. *Psychological Review*, *85*, 207–238.
Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
Nosofsky, R. M. (1988). Similarity, frequency and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 54–65.
Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 218–247). Oxford: Oxford University Press.
Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266–300.
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.
Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*, 53–79.
Palmeri, T. J. (1997). Exemplar similarity and the development of automaticity. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *23*, 324–354.
Payne, J. W., Bettman, J. R., & Johnson E. J. (1990). The adaptive decision maker. Effort and accuracy in choice. In R. M. Hogarth (Ed.), *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 129–153). Chicago: University of Chicago Press.
Persson, M. (1996). *Flexible and accurate reasoning with an exemplar-based algorithm: An examination of Gigerenzer and Goldstein (1996), based on a more realistic learning model*. Unpublished undergraduate thesis, Uppsala University, Department of Psychology.
Persson, M., & Juslin, P. (2000). Fast and frugal use of cue directions with few training exemplars. In L. Gleitman & A. K. Joshi (Eds.), *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum.
Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*, 29–46.
Ramsey, F. P. (1931). *The foundations of mathematics and other logical essays*. London: Routledge & Kegan.
Reisbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*. Mahwah, NJ: Lawrence Erlbaum.
Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization*. Hillsdale, NJ: Erlbaum.
Salmon, W. C. (1979). *The foundations of scientific inference*. Pittsburgh: Pittsburgh University Press.
Sieck, W. R., & Yates, J. F. (2001). Overconfidence effects in category learning: A comparison of connectionist and exemplar memory models. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1003–1021.
Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–19.
Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*, 3–21.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search for a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, *65*, 117–137.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*, 31–48.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.