WP : 112 ISSN : 1749-3641 (online)

The Similarity Heuristic

Daniel Read
Durham Business School
Mill Hill Lane, Durham DH1 3LB, United Kingdom
daniel.Read2@durham.ac.uk
Tel: +44 19 1334 5454 Fax: +44 19 1334 5201

Yael Grushka-Cockayne
London Business School
Regent's Park, London NW1 4SA, United Kingdom
ygrushka.phd2003@london.edu
Tel: +44 20 7000 8837 Fax: +44 20 7000 7001

The Similarity Heuristic

Decision makers are often called on to make snap judgments using fast-and- frugal decision rules called cognitive heuristics. Although early research into cognitive heuristics emphasized their limitations, more recent research has focused on their high level of accuracy. In this paper we investigate the performance a subset of the representativeness heuristic which we call the similarity heuristic. Decision makers who use it judge the likelihood that an instance is a member of one category rather than another by the degree to which it is similar to others in that category. We provide a mathematical model of the heuristic and test it experimentally in a trinomial environment. The similarity heuristic turns out to be a reliable and accurate choice rule and both choice and response time data suggest it is also how choices are made.

Keywords: heuristics and biases, fast-and-frugal heuristics, similarity, representative design, base-rate neglect, Bayesian inference

A heuristic is a decision rule that provides an approximate solution to a problem that either cannot be solved analytically or can only be solved at a great cost (Rozoff, 1964). Cognitive heuristics are analogous 'mental shortcuts' for making choices and judgments. Two familiar examples are the availability heuristic (judge an event frequency by the ease with which instances of the event can be recalled; Kahneman and Tversky, 1973), and the recognition heuristic (if you recognize only one item in a set, choose that one; Goldstein and Gigerenzer, 2002). Cognitive heuristics work by means of what Kahneman and Frederick (2002) call attribute substitution, by which a difficult or impossible judgment of one kind is substituted with a related and easier judgment of another kind. The recognition heuristic, for instance, substitutes the recognition of only a single option in a pair for the more costly process of searching for, selecting and evaluating information about both options. A central feature of cognitive heuristics is that while they are efficient in terms of time and processing resources, they achieve this at some cost in accuracy or generality. As an example, when events are highly memorable for reasons unrelated to frequency, the availability heuristic can overestimate their probability.

Early research into cognitive heuristics emphasized how they could produce systematic biases (Kahneman, Slovic & Tversky, 1982). Indeed, these biases were often the primary evidence that the heuristic was being used. Later research has emphasized the adaptive nature of heuristics, emphasizing their capacity to quickly and efficiently produce accurate inferences and judgments (Gigerenzer & Todd and the ABC research group, 1999; Samuels, Stich & Bishop, 2002). To use the term introduced by Gigerenzer and Goldstein (1996), heuristics are 'fast-and-frugal': they allow accurate decisions to be made quickly using relatively little information and processing capacity.

As Gilovich and Griffin (2003) observe, however, this new emphasis has not been applied to the 'classic' heuristics first described by Kahneman and Tversky (1973). One reason is that the two approaches to heuristics come from different research traditions that have asked different questions, and adopted correspondingly different methods. The modal question asked by the earliest researchers was 'do people use heuristic X?', while those in the

fast-and-frugal tradition started with 'how good is heuristic X?'. These two questions are answered using different research strategies. The first strategy is a form of what Brunswik (1955) called a *systematic* design, the second related to what he called a *representative* design. In a systematic design the stimuli are chosen to permit the efficient testing of hypotheses; in the representative design the stimuli are literally a *representative* sample, in the statistical sense, drawn from the domain to which the results are to be generalized (Dhami, Hertwig & Hoffrage, 2004).

If misinterpreted, the use of a systematic design can exaggerate the importance of atypical circumstances. The experimental conditions tested are usually chosen so that different judgment or choice rules predict different outcomes, and since one of those rules is usually the normatively optimal rule, and the purpose of the experiment is to show that a different rule is in operation, the experiment invariably reveals behavior that deviates from the normative rule. For instance, studies of the availability heuristic are designed to show that, whenever using the heuristic will lead to systematic under- or over-estimation of event frequency, this is what occurs. Many early observers concluded that such findings showed evidence of systematic and almost pathological *irrationality* (e.g. Nisbett & Ross, 1980; Piatelli-Palmarini, 1996; Plous, 1993; Sutherland, 1992). The extent of the irrationality observed, however, may have been the result of the use of a systematic design, combined with an interpretation of the results from using that design as being typical[i]. If the goal is to measure how well a decision rule or heuristic performs, a more representative design should be used[ii].

In this paper we investigate the *representativeness heuristic*, one of the classic heuristics first described by Kahneman and Tversky (1972), who defined it as follows:

> A person who follows this heuristic evaluates the probability of an uncertain event, or a sample, by the degree to which it is: [i] similar in essential properties to its parent population; and [ii] reflects the salient features of the process by which it is generated. (Kahneman & Tversky, 1972 p. 431)

The heuristic has two parts, one based on the similarity between sample and population, the other based on beliefs about the sampling process itself (Joram & Read, 1996). The focus in this paper is on one aspect of Part [i], which we refer to as the similarity heuristic[iii], according to which the judged similarity between an event and possible populations of events is substituted for its posterior probability. An example of this substitution is found in responses to the familiar "Linda" problem (Tversky & Kahneman, 1982). Because Linda is more similar to a 'feminist bank-teller' than a mere 'bank-teller,' she is judged to be more likely to be a feminist bank-teller (Shafir, Smith and Osherson, 1990).

An important study, using a systematic design, of what we call the similarity heuristic was conducted by Bar-Hillel (1974). Her subjects made judgments about sets of three bar charts like those in Figure 1, labeled *L*, *M* and *R* for left, middle and right. The *Similarity* group judged whether *M* was more similar to *L* or *R*. The *Likelihood of populations* group was told that *M* represented a sample that might have been drawn either from population *L* or *R*, and judged which population *M* was more likely to come from, and the *Likelihood of samples* group was told that *M* represented a population that might have generated either sample *L* or *R*, and judged which sample was more likely to be generated from *M*. If the similarity heuristic is used, all three judgments would coincide. Bar-Hillel systematically designed the materials so that this coincidence could easily be observed. All the triples had the following properties:

1. Every bar in *M* was midway in height between the bars of the same color in *L* and *R*.
2. The rank-order of the bar heights in *M* coincided with those in either *L* or *R*, but not both.
3. When *M* was interpreted as describing a population and *L* and *R* were interpreted as samples, then the sample with same rank-order as *M* was the least probable.
4. Likewise, when *L* and *R* were interpreted as populations and *M* as a sample, then *M* was less likely to be drawn from the population whose rank-order it matched.

In other words, the stimuli were systematically designed to ensure that, under both interpretations of likelihood, the objective odds favored the same chart, which was not the

chart with the same rank-order as *M*. In Figure 1, sample *M* is more likely to be drawn from population *R*, and sample *R* is more likely to be drawn from population *M*, although the rank-order of the bar-heights in *M* is the same as that of *L*. Bar-Hillel correctly anticipated that both similarity and likelihood judgments would be strongly influenced by rank-order.

—Figure 1 about here –

Although this study is very elegant, for our purposes it has two shortcomings, both related to the fact that the stimuli were highly unusual[iv]. First, the stimuli all had the same atypical pattern, which may have suggested the use of judgment rules that would not have been used otherwise. For instance, the rule 'choose the one with the same rank-order' was easy to derive from the stimuli, and could then be applied to every case – in other words, the attribute 'rank-order' rather than 'similarity' could have been substituted for 'likelihood.' This possibility is enhanced by the presentation of stimuli as bar charts rather than as disaggregated samples, and the use of lines to connect the bars. Both features make rank-order extremely salient.

Moreover, the use of a systematic design means the study does not indicate how accurate the similarity heuristic is relative to the optimal decision rule, even for bar charts connected by lines. When the majority similarity judgment is used to predict the majority choice in the *Likelihood of Populations* group, the error rate was *90%*. But since only a tiny proportion of cases actually meet the four conditions specified above, this number is practically unrelated to the overall accuracy of the heuristic. Indeed, the fact that respondents make errors in Bar-Hillel's study is highly dependent on the precise choice of stimuli. In the illustrative stimuli of Figure 1, if the bar heights in *L* are slightly changed to those indicated by the dashed lines (a 5% shift from yellow to green), then the correct answer changes from *L* to *R* (the probability that *R* is correct changes from .41 to .65).

One goal of the experiment described in this paper is to address the issues implied by this analysis. First, we elicit choices and judgments of similarity in an environment in which the relationship between sample and population varies randomly. Second, because we examine a random sample of patterns in this environment, we are able to assess the efficiency of the similarity heuristic. Our method was deliberately designed to find a point of contact between the two traditions of research in heuristics – the early tradition exemplified by Kahneman and Tversky's work, and the later tradition exemplified by the work of Gigerenzer and Goldstein (1996). Our research shows there is no fundamental divide between these traditions. As a first step, we describe a precise and testable model of the similarity heuristic.

A Model of the Similarity Heuristic

The similarity heuristic is a member of what is perhaps the broadest class of decision rules, those in which the decision to act on (or to choose, or to guess) one hypothesis rather than another is based on the relative value of a decision statistic computed for each hypothesis. In the most basic version of this class, one hypothesis is chosen because the decision statistic favors that hypothesis more than any other and, if two or more hypotheses share the same maximum decision statistic, one is chosen using a tie-breaking procedure. In the context of such models, a wide range of decision statistics have been proposed. Some of these are objective relationships between the data and the hypotheses. Amongst these are the likelihood, and the posterior probability computed from Bayes' rule. These decision statistics are particularly important because they constitute the theoretical benchmark for the performance of a decision rule. Several other "objective" decision statistics are those discussed recently by Nilsson, Olsson and Juslin (2005) in the context of probability judgment. Indeed, two of these are operationalizations of 'similarity' based on Medin and Schaffer's (1978) context theory of learning, comprising an adaptation of one interpretation of the representativeness heuristic originating in Kahneman and Frederick (2002), and the other is their own exemplar-based model. The decision statistic can also be – and indeed when

making choices typically is – a subjective relationship between data and hypothesis. Recognition is such a subjective relationship, where the recognition of an object can be used as the basis for making a judgment such as 'the object is large.' The feeling or judgment of similarity between data and hypothesis is another subjective relationship, and the one we focus on.

We will illustrate with a simple decision problem. Imagine you are birdwatching in a marshy area in South England, and hear a song that might belong to the redshank, a rare bird whose song can be confused with that of a common greenshank. You must decide whether or not to wade into the marsh in hope of seeing a redshank. In normative terms, your problem is whether the expected utility of searching (s) for the redshank is greater than that of not searching ($\bar{s}$):

$$p(r/d)u(s/r)+p(g/d)u(s/g)>p(r/d)u(\bar{s}/r)+p(g/d)u(\bar{s}/g), \qquad (1)$$

where $p(r/d)$ is the probability it is a redshank given the data (i.e., the song), $p(g/d)$ is the probability it is a greenshank given the data, $u(s/r)$ is the utility of searching given that it is a redshank, and so on. The probabilities are evaluated with Bayes' rule, which draws on likelihoods and the prior probability of each hypothesis, $p(r)$ and $p(g)$. If we substitute the multiplication *posterior* = *prior* $\times$ *likelihood* into (1), and rearrange terms, the decision rule is to search if

$$\frac{p(r)\,p(d/r)}{p(g)\,p(d/g)}>\frac{u(\bar{s}/g)-u(s/g)}{u(s/r)-u(\bar{s}/r)} \qquad (2)$$

If all the utilities are equal, this reduces to searching if $p(r)\,p(d/r)>p(g)\,p(d/g)$.

When using the similarity heuristic, the probabilities are replaced with similarity judgments, $s(d,r)$ and $s(d,g)$: respectively, the similarity of the song to the redshank's and the greenshank's. According to the similarity heuristic, you should search if

$$s(d,r)>s(d,g). \qquad (3)$$

That is, search if the birdsong you have just heard sounds (to you) more similar to that of the redshank than that of the greenshank.

Within a given environment, the theoretical performance of a decision rule can be estimated by computing the proportion of times it yields the correct answer, relative to the same proportion for the optimal decision rule. We show how to estimate the performance of the similarity heuristic against the Bayesian benchmark.

The decision model begins with a vector of decision statistics. For the similarity heuristic, these statistics are judgments of similarity between the sample or case (the data) and the population from which it might have been drawn[v]. For each of the *n* possible hypotheses, h_i, $i = 1,...,n,$ and the data, d_j, the decision maker generates a similarity judgment $s(d_j, h_i)$. The set of *n* judgments form a similarity vector $\mathbf{s}_j' = [s_{1j}, s_{2j}, ..., s_{nj}]$, where $s_{ij} = s(d_j, h_i)$.

Given the similarity vector, the next step is to pick out the maximum value from this vector, which is done by assigning *1* if s_{ij} takes the maximum value within $\mathbf{s_j}$, and *0* otherwise, yielding the *maximum similarity* vector, with the same dimensions as $\mathbf{s_j}$:

$$\mathbf{ms}_\mathbf{j}' = [ms_{1j}, ms_{2j}....ms_{nj}], \text{ where } ms_{ij} = \begin{cases} 1 & if\ s_{ij} = \max(\mathbf{s_j}) \\ 0 & otherwise \end{cases} \qquad (4)$$

In the simplest decision rule, h_i is chosen if the maximum similarity vector contains only a single value of *1* in the *i*-th position. If there is more than one such value, meaning that more than one hypothesis ties for maximum decision statistic, each candidate hypothesis has an equal chance of being chosen. The operation of this rule is implemented in the *decision* vector $\mathbf{ds_j}$:

$$\mathbf{ds}_\mathbf{j}' = [ds_{1j}, ds_{2j}, ..., ds_{nj}], \text{ where } ds_{ij} = \frac{ms_{ij}}{\sum_{i=1,...,n} ms_{ij}}, \qquad (5)$$

The value of ds_{ij}, therefore, is the probability the choice rule will select hypothesis h_i.

To calculate the probability that, for a given piece of evidence, this choice rule will select the correct option, we pre-multiply the decision vector by the vector of corresponding posterior probabilities ($\mathbf{pl'_j}$) computed using Bayes' rule:

$$\mathbf{pl'_j} = [pl_{1j}, pl_{2j} .... pl_{nj}], \text{ where } pl_{ij} = p(h_i/d_j) = \frac{p(h_i)\,p(d_j|h_i)}{\sum_{i=1,...,n} p(h_i)\,p(d_j|h_i)} \qquad (6)$$

Hence, given a set of hypotheses $H = \{h_i,\ i = 1,...,n\}$, a choice rule $\mathbf{s_j}$, prior probabilities $\mathbf{p}$, and evidence d_j, the *accuracy* of the choice rule, meaning the probability of making a correct decision, is given by:

$$A(\mathbf{s_j}, H, \mathbf{p}, d_j) = \mathbf{pl_j} \cdot \mathbf{ds_j} = \sum_{i=1,...,n} pl_{ij} ds_{ij} \qquad (7)$$

Next, we determine the performance of the choice rule given this hypothesis set and all possible evidence that might occur. The evidence could be, for instance, every bird song that might be heard. If the evidence is discrete (e.g., we might hear one of a finite number, *m*, of possible sounds) the corresponding mean accuracy is:

$$A(\mathbf{S}, H, \mathbf{p}) = \sum_{j=1}^{m} pd_j \sum_{i=1}^{n} pl_{ij} ds_{ij}\ , \qquad (8)$$

where $\mathbf{S}$ is the $n \times m$ matrix representing the similarity of each piece of evidence to each hypothesis, and pd_j denotes the probability of obtaining evidence d_j.

Just as the evidence can vary, so can the prior probabilities associated with a given set of hypotheses. For instance, you might be in a situation where house sparrows are rare and Spanish sparrows are common, or the reverse. To obtain the mean accuracy of the decision rule we need to carry out the summation in Eq. (8) over the entire space of possible prior probability distributions:

$$A(\mathbf{S}, H) = E(Correct\,|\,\mathbf{S}, H) = \sum_{k=1}^{r} pp^k \sum_{j=1}^{m} pd_j^k \sum_{i=1}^{n} pl_{ij}^k ds_{ij}\ , \qquad (9)$$

where H is the hypothesis set. The superscript k is added to the probabilities of obtaining evidence d_j, and to the posterior probabilities, to indicate that their values assume a specific vector k of possible priors. The summation is carried out over the discrete set of prior

probability vectors, while multiplying by the probability of each prior probability vector, denoted by pp^k. Note that while the operation of the similarity heuristic (although not its performance) is independent of the distribution of prior probabilities, other rules need not be. To model Bayes' rule, for instance, ds_{ij} in Eq. (9) is replaced by pl_{ij}^k .

The above analysis focuses on deterministic choice rules. Although this is not the place to develop theories of stochastic choice, they can be modeled by means of Monte Carlo simulations of *A(S,**H**)* in which the vectors (e.g., **s´**, **ms´, ds**) are changed in the relevant fashion. The role of error, for instance, can be modeled by laying a noise distribution over the similarity vector (**s´**), bias by systematically changing some values of the same vector, and a trembling hand by random or even systematic changes to the decision vector (**ds**)[vi].

We illustrate our analysis and some of its implications with a simulation of the likelihood heuristic, for which likelihoods, $p(d/h_i)$, rather than similarity judgments, are the decision statistic. Likelihoods are often taken as a proxy for similarity (Villejoubert & Mandel, 2002; Nilsson, Olsson & Juslin, 2005) and the representativeness heuristic has even been interpreted as being equivalent to the likelihood heuristic (Gigerenzer & Murray, 1987) This analysis, therefore, can provide us with some expectations about when the similarity heuristic is likely to perform well, and when it will perform poorly.

We consider a simple "binomial balls in urns" environment, such as the one adopted by Grether (1980, 1992) and Camerer (1987). Imagine two urns (the hypotheses), denoted *A* and *B*, each containing red and white balls in known proportions, denoted R_A and R_B, that is, $H = \{R_A, R_B\}$. The decision maker obtains a random sample of 5 balls drawn from an unseen urn, and must then bet on whether it was drawn from urn *A* or *B*. Corresponding to each possible sample, e.g., $d_j = \{RRWRW\}$, and each hypothesis, there is a likelihood, $l_{ij} = p(d_j/h_i)$, which can be computed from the binomial distribution. The decision statistic vector is the vector of likelihoods $\mathbf{l}'_\mathbf{j} = [l_{Aj}, l_{Bj}]$. Each such vector is

transformed, by means of Eq. (4) and (5), into a decision vector $\mathbf{dl}'_{\mathbf{j}}$, equal to $[1 \quad 0]$ if $l_{Aj} > l_{Bj}$, $\left[\frac{1}{2} \quad \frac{1}{2}\right]$ if $l_{Aj} = l_{Bj}$, and $[0 \quad 1]$ if $l_{Aj} < l_{Bj}$. The probability of a correct choice is obtained by pre-multiplying this decision vector by the posterior (Bayesian) probability vector, to give $A(\mathbf{l}_{\mathbf{j}}, H, \mathbf{p}, d_j)$. The overall accuracy of the likelihood heuristic, $A(\mathbf{L}, H, \mathbf{p})$, is obtained by computing the probability of correct choices for each sample, weighting each of these probabilities by the probability of obtaining the sample, and then summing these weighted probabilities.

Table 1 shows the results of this analysis. The top row shows hypothesis sets, chosen to represent a wide range of differences between populations. When $H = \{.5,.5\}$ the populations have no distinguishing characteristics, while when $H = \{.9,.1\}$ they look very different. In the identification of birds, a population of house sparrows and Spanish sparrows is close to the first case, while house sparrows and sparrow hawks are like the second. The first column in the table gives the prior probabilities for each urn, $\mathbf{p}' = [p_A, p_B]$. The final row in the table presents $A(\mathbf{L}, H, \mathbf{p})$, the average accuracy of the likelihood heuristic for each hypothesis set. Because the likelihood heuristic, like the similarity heuristic, is not influenced by prior probabilities this value is the same for all cells in its column. The values in the middle cells show the incremental accuracy from using Bayes' rule instead of the likelihood heuristic, given each vector of priors, i.e. $A(\mathbf{B}, H, \mathbf{p}) - A(\mathbf{L}, H, \mathbf{p})$.

If the likelihood heuristic is a good proxy for the similarity heuristic, this analysis indicates when the similarity heuristic is likely to perform well relative to Bayes' rule, and when it will perform poorly. These conditions were described formally by Edwards, Lindman & Savage (1963). Roughly, they are that (a) the likelihoods strongly favor some set of hypotheses; (b) the prior probabilities of these hypotheses are approximately equal; and (c) the prior probabilities of other hypotheses never 'enormously' exceed the average value in (b). In Table 1, condition (a) becomes increasingly applicable when moving from left to

right, and condition (b) when moving from bottom to top[vii]. If we replace 'likelihood' in (a) with 'similarity', then these are also the conditions in which the similarity heuristic is likely to perform well. Likewise, when the conditions are not met, the similarity heuristic will do poorly.

-- Table 1 about here –

The Experiment

We investigated how well the similarity heuristic performs as a choice rule, and whether people actually use it. In four experimental conditions, judgments or choices were made about two populations and a single sample. Separate groups assessed the similarity of the sample to the populations (a single estimate of $s(d,h_2)-s(d,h_1)$), or chose the population from which the sample was most likely to have been drawn.

The populations and samples were, like those in Bar-Hillel's (1974) study, drawn from a trinomial environment. Within this environment, we adopted a representative design. Two populations (hypotheses) were generated using a random sampling procedure. The populations used were the first 240 drawn using this procedure, which were randomly paired with one another. A random sample was then drawn, with replacement, from one of the populations in the pair, and the first sample drawn from each pair was the one used in the experiment. The populations and samples were shown as separate elements arranged in random order, as shown in Figure 2, and not in the form of summary statistics. We call each set of populations and sample a triple.

-- Figure 2 about here --

We also considered the relationship between and the similarity heuristic and the use of prior probability information. As discussed in section 2 above, the similarity heuristic makes

the same choice as Bayes' rule whenever $\text{sgn}\left[s(d,h_1)-s(d,h_2)\right]=\text{sgn}\left[p(h_1)p(d/h_1)-p(h_2)p(d/h_2)\right]$. Since the similarity heuristic disregards prior probabilities, it can lead to error when these are not equal $\left(p(h_1)\neq p(h_2)\right)$. In the experiment we chose the population from which the sample was chosen with a (virtual) throw of the dice, with prior probabilities of 1/6 and 5/6. One choice group had knowledge of the prior probabilities, while another group did not.

Method

Subjects

We tested 160 participants, all members of the London School of Economics community. In return for their participation, respondents received a £2 ($4) voucher for Starbucks.

Materials

The materials were based on 120 triples, each comprising two populations and one sample of red, yellow and blue rectangles. The population generating algorithm was as follows. First, we chose a number between 0 and 100 from a uniform distribution and specified this as the number of blue rectangles (call it *b*); next, we generated a number between 0 and (100-*b*) from a uniform distribution, and specified this as the number of green rectangles (*g*). The number of yellow rectangles was therefore *y=100-b-g*. This yielded populations with an average of 50 blue, 25 green and 25 yellow rectangles. In this way we generated populations that were, on average, composed of a large number of blue rectangles. This is analogous to many natural populations, in which the modal member is of one type, but in which alternative types are also relatively abundant – such as the ethnic composition of European and North American cities, or bird populations pretty well everywhere.

For each question, we randomly generated a pair of populations, one of which was assigned a high prior of 5/6, the other a low prior of 1/6. One population was chosen with probability equal to its prior, and a sample of 25 rectangles was drawn (with replacement)

from this population. We used the first 120 stimuli sets generated, and they were presented in the order generated.

Procedure

Each respondent made judgments or choices for 30 triples, so the 120 triples comprised four replications of the basic design. Within each replication, there were 10 participants in each of four groups: The *Similarity* group were told nothing about the context, and simply rated which of the larger sets of rectangles the small set was more similar to; the *Similarity/Population* group made similarity judgments, this time with full knowledge that the sets represented two populations and one sample; the *Choice/No prior* group guessed which population the sample came from without knowledge of prior probabilities; and the *Choice/Prior* group made the same choice but with this knowledge.

In all conditions, respondents first read an introductory screen which told them they would be asked questions about 'sets of rectangles' and were shown an unlabelled example of such sets. The instructions then diverged, depending on the experimental condition. Those in the *Similarity* group read *You will see two large sets and one small set like the following* and were shown a triple like that in Figure 2, with the three sets labeled, respectively, as *Large Set 1*, *Small Set* and *Large Set 2*. For each subsequent triple, they indicated which large set the small set was more similar to, using a 9-point scale that ranged from *Much more similar to LS 1* to *Much more similar to LS 2*.

The instructions for the remaining groups included the following description of the task context:

> *We want you to consider the following procedure. First, we randomly generated two populations of yellow, red and blue rectangles, which we call Population 1 and Population 2.* [Here the *Choice/Prior* group received information about prior probabilities, as described later…]

Then we drew a sample of 25 rectangles from either Population 1 or Population 2. [Here an example was shown, with the sets labeled as *Population 1*, *Sample*, and *Population 2*.]

We drew the sample this way:

We randomly drew one rectangle and noted its color.

Then, we returned the rectangle to the population and drew another one, until we had drawn 25 rectangles.

The sample could have been drawn from either Population 1 or Population 2.

Those in the *Similarity/Population* group then judged the similarity of the sample to Population 1 or Population 2 using the 9-point scale, this time with the endpoints labeled *Much more similar to Population 1* and *Much more similar to Population 2*.

For those in the two choice groups the task was to indicate which population they thought the sample came from. This was done by clicking one of two radio keys. The instructions for the *Choice/Prior* group included the following information:

First [… as above].

Second, we rolled a die. If any number from 1 to 5 came up, we drew a sample of 25 rectangles from one population, while if the number 6 came up, we drew a sample of 25 rectangles from the other population.

In the following example we drew a sample from Population 1 if the numbers 1 to 5 came up, and drew a sample from Population 2 in the number 6 came up. [Here an example was shown, with five dice faces above Population 1, and one above Population 2.] *In the following example we drew a sample from Population 2 if the numbers 1 to 5 came up, and drew a sample from Population 1 if the number 6 came up.* [Here the example had one face above Population 1 and five above Population 2].

Once the population was chosen, we drew the sample this way […] the standard instructions followed, ending with …] *The sample could have been drawn from either Population 1 or Population 2, depending on the roll of the die.*

For each triple in the *Choice/prior* group five dice faces were above the high prior population and one face above the low prior population. The population number of the high prior population was randomized.

In all conditions we recorded the time taken to make a choice or similarity judgment.

Results

How reliable and consistent are judgments of similarity?

For similarity to be a reliable and valid basis for making probabilistic choices, there must be some "common core" underlying the similarity judgments made by different people and in different contexts. We measured this core by evaluating the inter-context and inter-subject consistency of similarity judgments. There were four sets of 30 triples, each of which received similarity judgments from 20 subjects, 10 each from the *Similarity* and *Similarity/Population* groups. For each set of triples, we computed the mean inter-subject correlation, both within and between experimental groups. These are shown in Table 2. As can be seen, the mean inter-subject correlation was high (overall ranging from .71 to .79) and there was no appreciable reduction in this value when attention was restricted to correlations between subjects in different groups (ranging from .68 to .79).

-- Table 2 about here –

Given the high correlation between individual judgments, it is not surprising that the correlation between the *average* similarity judgments for the 120 questions was extremely high (.95). Moreover, even the mean similarity judgments in the two groups were almost identical (5.06 vs 5.05), indicating that in both conditions the scale was used in the same way. Finally, to anticipate the next section, the proportion of correct choices predicted by both measures of similarity was almost identical. We conducted two logistic regressions, using similarity ratings to predict the optimal Bayesian choice (we will call this *BayesChoice*). The percentage of correct predictions was 86% for both Similarity groups, and these were

distributed almost identically across both Populations 1 and 2. Because the two similarity measures are statistically interchangeable, we usually report results from combining the two measures.

Overall, these analyses show that the judgments of similarity in both contexts contained a substantial common core. We conclude, therefore, that similarity judgments are reliable. We next turn to the question of their validity as a basis for probabilistic choice.

How accurate is the similarity heuristic?

We simulated the performance of the similarity heuristic in two ways. First, we examined the correlation between the 9-point similarity rating and the option that would be chosen by an optimal application of Bayes' rule (denoted *BayesChoice*). Figure 3 shows the proportion of times *BayesChoice* equals Population 2, for each level of *Similarity*. This proportion increases monotonically in an *S*-shaped pattern, with virtually no *Population 2* options predicted when *Similarity=1* and almost 100% when *Similarity=9*. The correlation between *individual* similarity judgments and *BayesChoice* is .76.

—Figure 3 about here --

We also compared the accuracy of the similarity heuristic with that achieved using Bayes' rule and the likelihood heuristic (*BayesChoice* and *LKChoice*). We simulated the heuristic using the principles described previously: if the *Similarity* rating was less than 5 (i.e., implying $s(d,h_1) > s(d,h_2)$) then predict a choice of Population 1, if it is equal to 5 then predict either population with probability of .5, otherwise predict Population 2 (we use *SimChoice* to denote these individual simulated choices). *Simchoice* correctly predicted the population from which the sample was drawn 86% of the time, compared to 94% for *LKChoice* and 97% for *BayesChoice*.

Because similarity is a psychological judgment it is, unlike likelihoods and prior probabilities, prone to error. To obtain a low-error judgment of similarity, we took the mean

similarity judgment for each question and applied our decision rule to this mean (i.e., if mean *Similarity* < 5 choose Population 1, etc.). We denote these choices *Simchoice/A* (for aggregate). Relative to *Simchoice*, using *Simchoice/A* increased the correlation between the similarity heuristic and *BayesChoice* from .76 to .85, and increased overall accuracy from 86% to 92%.

In this context, therefore, the similarity heuristic achieves a high level of accuracy when making probabilistic choices. But this does not demonstrate that people actually take the opportunity to use similarity when making choices. This is what we evaluate next.

Do people use the similarity heuristic?

Similarity/Choice agreement. For each respondent in the two choice groups, we compared the choices they made to the predictions of *Simchoice/A*. Figure 4 shows, for each respondent in the *Choice/No prior* and *Choice/Prior* groups, the proportion of correct predictions. There was an extremely good fit between actual and predicted choices: an average of 89% predictions in the *No prior* group (Median 92%), and 86% in the *Prior* group (Median 90%).

—Figure 4 about here --

This is not an irrefutable demonstration that people use the similarity heuristic, since both choice and similarity judgments are also highly correlated with *BayesChoice*, leaving open the possibility that the similarity/choice relationship might not be causal (i.e., similarity determines choice), but merely due to the use of another choice rule (or rules) that is correlated with both similarity and Bayes rule. We therefore conducted two additional analyses to consider whether the similarity heuristic predicts choice beyond that predicted by *BayesChoice*. First, we conducted a logistic regression in which individual choices (in both the *Choice/No prior* and *Choice/Prior* conditions) was regressed on the mean *Similarity*

rating, the normalized likelihood ratio (*NLKR*) defined as $\frac{p(d/h_2)}{1+p(d/h_2)}$, and the prior probability of Population 2. The model was chosen using a forward selection procedure (probability for entry = .10., for removal = .15). In both analyses, mean *Similarity* was the most significant predictor in the final model. The logits (log odds) for the final models were:

Choice/No-prior: *4.03 – 0.63 Similarity – 2.32 NLKR*

Choice/Prior: *5.51 – 0.89 Similarity – 2.10 Prior*

All coefficients were highly significant (p-value for Wald statistic < .0001), and classification accuracy was 88% for the *No prior* group and 87% for the *Prior* group. This is strong evidence that the similarity heuristic was being used by both groups. Separate regressions including only *Similarity* as an explanatory variable supported this view – classification accuracy was reduced by less than 1% in both groups.

Finally, to provide the strongest possible test we conduct a further analysis relating *individual* similarity judgments to *individual* choices. Because we did not collect similarity judgments and choices from the same respondents, we created "quasi-subjects," simply by placing the individual responses in all four conditions into four columns of our data file, and then analyzing the relationships between conditions as if they had been collected from the same respondent. We lined up, for instance, the response from the first respondent who made a similarity judgment to one item, with the first respondent who made a choice to that item, and so forth. Our reasoning was that if the similarity heuristic is robust to being tested under these unpromising circumstances, it will surely be robust to tests when both choices and similarity judgments come from the same respondent.

-- Table 3 about here --

We conducted two correlational analyses of these data, as shown in Table 3. First, we looked at the first order correlation between *Simchoice*, *Simchoice/Pop*, *Choice/Prior* and *Choice/No prior*. These were, as can be seen in Table 3, moderately high (≅ .6) and overwhelmingly significant. This indicates that the relationship found with the aggregate similarity judgments does not vanish when they are disaggregated. We then conducted the same analysis, but this time partialling out three alternate choice predictors: *LKChoice*, *BayesChoice*, and the *Prior* – these predictors are all highly intercorrelated but we included them to squeeze out the maximum predictive power. The partial correlations were reduced, but all remained positive and significant. Thus, individual similarity judgments made by one respondent were able to robustly predict the individual choices made by another respondent[viii].

Response times. A further line of evidence that choice is based on the similarity heuristic comes from the pattern of response times (*RT*s), which suggest that both choices and similarity judgments are driven by the same psychological process. Figure 5 is a boxplot showing the distribution of median *RTs* for each triple, for all four conditions. This shows the average *RT* and its distribution and its distribution, is approximately the same for all conditions, an observation supported by a non-significant ANOVA ($F(3,\ 357) = 1.7,\ p > .15$).

—Figure 5 about here –

Table 4 shows correlations between median *RT*s for all triples. All the relationships are highly significant ($p < .0001$, $n = 120$) and, more importantly, correlations within response categories (*Similarity* with *Similarity/Population*, and *Choice/No prior* with *Choice/Prior*, Mean r = .70) are close to those between categories (*Similarity* with *Choice*, Mean r=.65). This occurs despite an undoubted level of method variance due to the different response formats in the two categories (a choice between two radio keys versus rating on a 9-point scale).

-- Table 4 about here –

Moreover, choice response times show a relationship that should be expected if similarity judgments are the basis for choice. When the sample is equally similar to the two populations (i.e., similarity judgments are close to the scale midpoint) it also takes longer to choose which population it came from. Figure 6 plots the median response time for all 120 questions against the average Similarity judgment for each question, along with the best fitting quadratic function. In both cases this function revealed the expected significant inverted-U function[ix].

-- Figure 6 about here --

Overall, therefore, analysis of the responses made and the time taken to make them closely fit what we would expect if choices are based on the similarity heuristic.

How is prior probability information used?

Consistent with much earlier research (e.g., Gigerenzer, Hell & Blank, 1988; Fischhoff, Slovic & Lichtenstein, 1979), we found that prior probabilities influenced choice in the right direction but were underweighted. Respondents in the *Choice/Prior* condition were significantly more likely to choose the high prior item than were those in the *Choice/No Prior*

condition (76% versus 71%; $F(1, 119) = 20.4$, $\varepsilon^2 = .146$, $p < .001$), although they still chose it at a lower rate than the actual prior probability (83%, or 5/6). Our design enabled us to go further and determine whether knowledge of prior probabilities improved choice, and more generally whether the knowledge was used strategically.

Knowledge of priors did not increase accuracy, which was 86.3% in the *Choice/Prior* condition and 86.1% in the *Choice/No prior* condition ($F(1, 119) < 1$). This suggests that knowledge about prior probabilities was used inefficiently. This is illustrated in Figure 7, which shows, for both choice groups, the proportion of times the correct choice was made when the sample was drawn from high prior population versus when it was drawn from the low prior population (we will say, when the prior is *consistent* and *inconsistent*). When the prior was consistent, the *Choice/Prior* group was a little more accurate than the *Choice/No prior* group (90% versus 87%), but when it was inconsistent, they were much less accurate (74% versus 82%). This was reliable result: an ANOVA with the group as a within-triple factor, and consistency of priors as a between-triple factor, revealed a highly significant interaction, $F(1, 118) = 17.7$, $\varepsilon^2 = .131$, $p < .001$. Since the prior was consistent 83% of the time, the small benefit it gave when consistent was counterbalanced by the larger cost when it was inconsistent.

-- Figure 7 about here --

A strategic way to combine knowledge of prior probabilities with similarity data is to go with the high prior option when the sample is equally similar to both populations, but to go with similarity when it is highly similar to only one population. This can be seen by referring to Table 1: knowledge of priors is less useful when the environment is represented by the columns to the right, when the two hypotheses are highly distinguishable, than when it is represented by the columns to the left. We investigated to what degree respondents were strategically putting more weight on priors when they found themselves in situations like the

left rather than the right columns. The fact that performance was not improved by knowledge of priors suggests they were not using the information strategically, and we confirmed this by examining the difference between the proportion of time the high prior item was chosen in the *Choice/Prior* versus *Choice/No prior* groups, as a function of similarity judgments. We define *PrEqHi* and *NoPrEqHi* as, respectively, the proportion of times the *Choice/Prior* and *Choice/No prior* groups chose the high prior option for each triple, and then computed a *proportional shift statistic* (*PSS*) for each triple, which was an index of the increase in choices of the high prior item in response to having that information.

$$PSS_i = \begin{cases} \dfrac{PrEqHi - NoPrEqHi}{1 - NoPrEqHi} & if\ \ PrEqHi > NoPrEqHi \\ \dfrac{PrEqHi - NoPrEqHi}{1 - PrEqHi} & if\ \ PrEqHi \le NoPrEqHi \end{cases}$$

The subscript *i* indexes the triple. *PSS* ranges from -1 to 1, the difference between the proportion of choices of the high prior option in the two choice conditions, divided by the maximum possible proportion of such choices. For example, if for one triple 90% of the *Choice/Prior* group chose the high prior item, as opposed to 80% of the *Choice/No prior* group, then PSS_i would be $\frac{.9-.8}{1.0-.8}=.5$. On the other hand, if 90% in the *Choice/No prior* group chose the high prior item while only 80% in the *Choice/Prior* group did, then PSS_i =(-.5). Because *PSS* cannot be computed if both *PrEqHi* and *NoPrEqHi* are equal to 1, which occurred in 33 cases, we obtained 87 usable values of *PSS*, with a mean value of .13 (*SD*=.62). The fact that the number is positive indicates respondents were more likely to choose the high prior item when they knew which one it was, and the specific value obtained can be interpreted as follows: for the average triple, if the high prior item was chosen by a proportion *p* of those in the *Choice/No prior* group, then it was chosen by $p+.13(1-p)$ of those in the *Choice/Prior* group.

Figure 8 shows the 87 values of *PSS* as a function of the mean similarity rating for each triple, along with the best fitting quadratic function. If knowledge of prior probabilities was being used strategically, this best-fitting function would have an inverse-U shape, indicating that prior probabilities had their greatest influence when the sample was equally similar to both populations. In fact, the quadratic function has the opposite shape to this hypothesized inverse-U, although it accounts for relatively little of the variance in *PSS* (R^2=*.021*). That is, while knowledge of population prior probability did increase the tendency to choose the high prior item, it did so indiscriminately – respondents in the Choice/Prior condition put equal weight on the prior when similarity was undiagnostic (when knowledge of the prior would be useful) than when it was diagnostic (and the knowledge was relatively useless).

—Figure 8 about here –

Discussion

Willard Quine famously described the problem of induction as being a question about the use of what we call the similarity heuristic:

> For me, then, the problem of induction is a problem about the world: a problem of how we, as we now are (by our present scientific lights), in a world we never made, should stand better than random or coin-tossing chances of coming out right, when we predict by inductions which are based on our innate, scientifically unjustified similarity standard. (Quine, 1969, p. 127).

Our research can be viewed as an investigation into just how much better than 'random' are these predictions, and our findings are that they are, at least in one context, very much better. In the environment in which our respondents found themselves, individual similarity judgments were able to come out right 86% of the time, compared to coin-tossing chances of 50%. Moreover, we also found strong evidence that people were using a *shared*, if not

necessarily innate, similarity standard to make their choices – the similarity judgments made by one group proved to be an excellent predictor of both the similarity judgments and the choices made by other groups.

As we noted earlier, although the similarity heuristic is a subset of the representativeness heuristic first described by Kahneman and Tversky (1972), we modeled our approach on the program of a different school of researchers. This program, well-summarized in Goldstein and Gigerenzer's (2002) seminal article on the recognition heuristic, is to:

> design and test computational models of [cognitive] heuristics that are (a) ecologically rational (i.e., they exploit structures of information in the environment), (b) founded in evolved psychological capacities such as memory and the perceptual system, (c) fast, frugal and simple [and *accurate*] enough to operate effectively when time, knowledge and computational might are limited, (d) precise enough to be modeled computationally, and (e) powerful enough to model both good and poor reasoning. (p.75)

In the rest of this discussion we comment on the relationship between this program and our own investigations.

Ecological rationality

The concept of ecological rationality is best described by the means of the lens model of Brunswik (1952, 1955; c.f. Dhami et. al, 2004), a familiar modernized version of which is shown in Figure 9 (e.g., Hammond, 1996). The judge or decision maker seeks to evaluate an unobservable criterion, such as a magnitude or probability. While she cannot observe the criterion directly, she can observe one or more fallible cues or indicators (denoted *I* in the figure) that are correlated with the criterion. Judgments are based on the observable indicators, and the accuracy (or 'ecological rationality') of those judgments is indexed by their correlation with the unobservable variable. For the recognition heuristic, the judgment is recognition ("I have seen this before"), which is a valid predictor of many otherwise

unobservable criteria (e.g., size of cities, company earnings), because it is itself causally linked to numerous indicators of those criteria (e.g., appearance in newspapers or on TV).

-- Figure 9 about here –

The ecological rationality of the similarity heuristic arises for similar reasons. Although researchers do not yet have a complete understanding of how similarity judgments are made, we do know that the similarity between a case *x* and another case or class *A* or *B* is a function of shared and distinctive features and characteristics (see Goldstone & Son, 2005, for a review). Likewise, the probability that *x* is a sample from a given population is closely related to the characteristics that *x* shares and does not share with other members of that population. It is perhaps not surprising, therefore, that similarity turns out to be such a reliable and valid index of class membership.

Evolved psychological capacities

Both the recognition and similarity heuristics work through a process of attribute substitution (recognition substituted for knowledge of magnitude, similarity substituted for knowledge of posterior probabilities), and are effective because of the strong correlation between the attribute being substituted for and its substitution. The reason for this high correlation is because both the capacity to recognize and the capacity to detect similarity are both products of natural selection.

The ability to assess the similarity between two objects, or between one object and the members of a class of objects, is central to any act of generalization (e.g., Attneave, 1950; Goldstone & Son, 2005). As Quine (1969) observed, to acquire even the simplest concept (such as 'yellow') requires 'a fully functioning sense of similarity, and relative similarity at that: *a* is more similar to *b* than to *c*' (p. 122). Some such 'sense of similarity' is undoubtedly innate. Children are observed making similarity judgments as early as it is possible to make the observations (e.g., Smith, 1989), and it is one of the 'automatic' cognitive processes that

remain when capacity is limited by time pressure or divided attention (Smith & Kemler-Nelson, 1984; Ward, 1983). Like recognition and recall, therefore, the ability to judge similarity is a skill we are born with and can deploy at minimal cognitive cost whenever it can serve our purposes. The similarity heuristic, like other fast-and-frugal heuristics, operates by 'piggy-backing' on this innate ability when probability judgments are to be made.

Although we have spoken blithely about 'similarity judgments' we recognize that these judgments are embedded in specific contexts. For instance, if asked to judge the similarity between a celery stick, a rhubarb stalk and an apple, the judgment *s(apple, rhubarb)* will be greater than *s(celery, rhubarb)* if the criterion is 'dessert' than if it is 'shape.' Indeed, the concept of similarity has been widely criticized because of this. Medin, Goldstone and Gentner (1993) give a concise summary of this critique:

> The only way to make similarity nonarbitrary is to constrain the predicates that apply or enter into the computation of similarity. It is these constraints and not some abstract principle of similarity that should enter one's accounts of induction, categorization, and problem solving. To gloss over the need to identify these constraints by appealing to similarity is to ignore the central issue. (p. 255).

This criticism is related to the question of whether the concept of similarity can be fully defined is a context free manner. It is likely that it cannot. The criticism does not, however, bear on the question of *whether* people make similarity judgments, nor whether those judgments are reliable. It is clear that people do and the judgments are. In our study, the correlation between average similarity judgments in different contexts was extremely high (.95), but this is not an isolated result – even in studies designed to distinguish between theories of similarity, similarity judgments are highly correlated across conditions. For instance, in a study using a systematic design to demonstrate asymmetry in similarity judgments, Medin et. al. (1993) obtained the expected asymmetries, yet the correlation between the average similarity judgments for the same pairs in different contexts was .91 (see their Table 1 for data; studies reported in Tversky and Gati, 1978, all yield the same conclusions). It appears that however people make their judgments of similarity these

judgments are (a) highly consistent across contexts and across people, (b) good predictors of the likelihood that a sample comes from a population, and (c) actually used to make these judgments of likelihood.

Fast, frugal, simple and accurate

These criteria concern the *relative* performance of heuristics. We can readily suggest ideal benchmarks for each criterion, but the standard that must be reached for us to say that the heuristic is frugal or fast or accurate is a matter for judgment and context. We will give an account of the performance of the similarity heuristic on some measures of these criteria, along with an indication of our own opinion about whether the heuristic reaches one standard or another.

When measuring the speed of a decision process, the optimum time is always 0 seconds. No actual process can achieve this, but the time taken to make a judgment of similarity was typically about 6 seconds (as shown in Figure 5). Although we cannot benchmark this time against other tasks, we suggest it is very little time given that it involved *two* similarity judgments, a comparison between them, and a physical response on a 9-point scale.

We can assess simplicity and frugality by comparing the similarity heuristic to the process of making judgments by means of Bayes' rule. A quantitative estimate can be derived by drawing on the concept of Elementary Information Process (EIP), introduced by Payne, Bettmann and Johnson (1993), to measure the effort required to perform a cognitive task. An EIP is a basic cognitive transformation or operation, such as making comparisons or adding numbers. Consider the simple case, as in our experiment, of a choice between two hypotheses given one piece of data. The similarity heuristic, as described in Eq. (3), requires three EIPs: two judgments of similarity, and one comparison between them. To apply Bayes' rule, in contrast, requires seven EIPs, as in the reduced form of Eq. (2): four calculations (two priors and two likelihoods), two products (multiplication of priors by likelihoods) and one comparison (between the products). Using this measure, Bayes' rule is more than twice as

costly as the similarity heuristic[x]. Moreover, not all EIPs are equal: if it is harder to multiply probabilities and likelihoods than to make ordinal comparisons, and harder to estimate likelihoods than to make judgments of similarity, then the advantage of the similarity heuristic grows. Clearly, the similarity heuristic is frugal relative to the Bayesian decision rule.

The similarity heuristic also performed much better than chance and proved to be a reliable choice rule. It is worth observing here that the location of one source of disagreement between researchers in the two heuristics 'traditions' is exemplified by the contrast between the accuracy achieved in our study, and that achieved by the earlier study of Bar-Hillel. Bar-Hillel (1974) observed accuracy of 10%, based on group data, while the corresponding value in our study is 92% (for group data, 86% for individual judgments). Moreover, this value of 92% is achieved despite the complicating factor of a prior probability not known to those making similarity judgments, and to a less transparent way of presenting information (as disaggregated populations and samples rather than graphs). The difference in studies is found in the choice of design. We drew on the ideals of the representative design described by Brunswik (1955), and argued for by Gigerenzer and Goldstein (1996). Once we established a random sampling procedure, we did not further constrain our samples to have any specific properties. Bar-Hillel (1974), on the other hand, deliberately chose items for which the theorized decision-rule and Bayes' rule would yield different choices. If we took Bar-Hillel's study as providing a test of the accuracy of the similarity heuristic, we would conclude that it was highly inaccurate. This would obviously be an illegitimate conclusion (and one that Bar-Hillel did *not* draw).

There is an additional methodological lesson to be drawn from a comparison between Bar-Hillel's (1974) study and ours. Although the normative performance of the similarity heuristic differed greatly between studies, the degree to which the heuristic predicted choice did not. Bar-Hillel reported her data in the form of a cross-tabulation between choices based on the average similarity judgment for each triple (in her case a two-point scale) and the majority choice for triples. In Table 5 we show her original data and compare it to the same analysis conducted for our data. The patterns of results are readily comparable, and lead to

the same conclusions not just about whether the similarity heuristic predicts choice, but even about the approximate strength of the relationship between choice and judgment.

-- Table 5 about here –

Precise enough to be modeled computationally

The similarity heuristic is also precise enough to be modeled computationally. In an earlier section we provided a general mathematical model of the similarity heuristic. It was not the only possible model; in fact, it was the *simplest* one. However, it turned out to be a very good model in the context of our experiment. When similarity judgments made by one group are used to predict the choices of another group, they predict those choices remarkably well.

Powerful enough to model both good and poor reasoning

All heuristics have a domain in which their application is appropriate, and when they step outside that domain they can go wrong. We have already considered the performance of the likelihood heuristic as a proxy for the similarity heuristic, and suggested the similarity heuristic will be most accurate when the likelihood heuristic is, and inaccurate when it is not. Specifically, and as shown formally by Edwards et al. (1963), the similarity heuristic can go wrong when some hypotheses have exceedingly low priors, and when the similarity judgments *s(d,h)* do not strongly differentiate between hypotheses.

A fascinating recent case in which the ideal conditions are *not* met, and the similarity heuristic (probably coupled with some wishful thinking) leads to some unlikely judgments is found in the scientific debate surrounding the identification of some observed woodpeckers, which might be of the ivory-billed or pileated species (White, 2006; Fitzpatrick et al, 2005). The two birds are *very* similar. Careful scrutiny can distinguish them, although to the untutored eye they would be practically identical. The prior probabilities of the two hypotheses, however, are not even remotely close to equal. The pileated woodpecker is

relatively common, but the last definite sighting of the ivory billed woodpecker was in 1944, and there is every reason to believe it is extinct (i.e., prior ≈ 0). It is interesting to observe, however, that the debate over whether some reported sightings of the ivory-billed woodpecker are genuine involves a 'scientific' application of the similarity heuristic (focusing on issues like the size of the bird and wing patterns), with little explicit reference to prior probabilities, even by skeptics[xi].

The ivory-billed woodpecker case is, however, uncharacteristic and understates the power of the similarity heuristic even when priors are extremely low. In the case of the ivory billed woodpecker, prior probabilities should play such a large role because of a conjunction of two factors: similarity is practically undiagnostic (only very enthusiastic observers can claim that the poor quality video evidence looks a *lot* more like an ivory-billed than pileated woodpecker), and the least-likely hypothesis has a very low prior probability. The situation is therefore like that in the bottom left-hand cell of Table 1.

But suppose the situation were different, and while the prior probability is very close to zero, similarity is very diagnostic. You are out strolling one day in a dry area a long way from water, an area in which you *know* there are no swans, which only live on or very near water. Yet you stumble across a bird that is very similar to a mute swan: It is a huge white bird with a black forehead and a long gracefully curved neck; its feet are webbed, it does not fly when you approach but raises its wings in a characteristic 'sail pattern' revealing a wingspan of about 1.5 meters. Even though the prior probability of seeing a swan in this location is roughly 0 (i.e., this is what you would say if someone asked you the probability that the next bird you saw would be a swan), you will not even momentarily entertain the possibility that this is one of the candidates having a very high prior (such as a crow, if you are in the English countryside). We suggest that most everyday cases are like the swan rather the woodpecker – similarity is overwhelmingly diagnostic, and is an excellent guide to choice and decision even in the face of most unpromising priors. This is why, to return to Quine, we can do so well using our 'innate, scientifically unjustified similarity standard.'

References

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology* 63 (4), 516-556.

Bar-Hillel, M. (1974). Similarity and probability. *Organizational Behavior and Human Performance* 11, 277-282.

Brunswik, E. (1952). The conceptual framework of psychology. *International encyclopedia of unified science*, 1, 656-760.

Brunswik, E. (1955). Symposium on the Probability Approach in Psychology: Representative design and Probabilistic Theory in a Functional Psychology. *Psychological Review* 62 (3), 193-217.

Camerer, C.F. (1987). Do Biases in Probability Judgment Matter in Markets? Experimental Evidence. *American Economic Review* 77 (5): 981-997.

Dhami, M. K., Hertwig, R. & Hoffrage, U. (2004). The role of representative design in an ecological approach to cognition. *Psychological Bulletin*, 130, 959-988.

Edwards, W.H., Lindman H. & Savage, L.J. (1963). Bayesian statistical inference for psychological research. *Psychological Review* 70 (3), 193-242.

Fischhoff B., Slovic P. & Lichtenstein S. (1979). Subjective sensitivity analysis. *Organizational Behavior and Human Performance* 23 (3), 339-359.

Fitzpatrick, J.W., Lammertink, M., Luneau Jr, M.D., Gallagher, T.W., Harrison, B.R., Sparling, G.M., Rosenberg, K.V., Rohrbaugh, R.W., Swarthout, E.C.H., Wrege, P.H., Swarthout, S.B., Dantzker, M.S., Charif, R.A., Barksdale, T.R., Remsen, J.V. Jn., Simon, S.D. & Zollner, D. (2005). Ivory-billed Woodpecker (Campephilus principalis) persists in Continental North America. *Science* 308, 1460-1462.

Gigerenzer, G. & Goldstein, D.G. (1996). Reasoning the Fast and Frugal Way: Models of Bounded Rationality. *Psychological Review* 103 (4), 650-669

Gigerenzer, G., Hell, W. & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance* 14, 513-525.

Gigerenzer, G. & Murray, D. (1987). *Cognition as intuitive statistics*. Hillsdale, NJ: Erlbaum.

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

Gilovich, T. & Griffin, D. (2003). Introduction – Heuristics and Biases: Then and Now. In Gilovich, T., Griffin, D. & Kahneman, D. (eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press.

Goldstein, D.G. & Gigerenzer, G. (2002). Model of Ecological Rationality: The Recognition Heuristic. *Psychological Review* 109 (1), 75-90

Goldstone, R. L., & Son, J. (2005). Similarity. In Holyoak, K.& Morrison, R. (Eds.). *Handbook of Thinking and Reasoning*. Cambridge, England: Cambridge University Press.

Grether, D.M. (1980). Bayes rule as a descriptive statistic: The representativeness heuristic. *Quarterly Journal of Economics* 95, 537-557.

Grether, D.M. (1992). Testing Bayes rule and the Representativeness heuristic: some experimental evidence. *Journal of Economic Behaviour and Organization* 17, 31-57.

Hammond, K. R. (1996). *Human judgment and social policy*. Oxford: Oxford University Press.

Joram, E. & Read, D. (1996). Two faces of representativeness: The effects of response format on beliefs about random sampling. *Journal of Behavioural Decision Making* 9, 249-264.

Kahneman, D. & Fredrick S. (2002). Representativeness Revisited: Attribute Substitution in Intuitive Judgment in Gilovich, T., Griffin, D. & Kahneman, D. (eds). (2003). *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press.

Kahneman, D., Slovic, P. & Tversky, A. (eds). (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

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

Kahneman, D. & Tversky, A. (1973). On the psychology of prediction. *Psychological Review* 80, 237-251.

Kemp, C., Bernstein, A. & Tenenbaum J.B. (2005). A generative theory of similarity. *Proceedings of the 27th Annual Conference of the Cognitive Science Society.*

Medin, D. L., Goldstone R. L. & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254-278.

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

Navarro, D. J. & Lee, M. D. (2004). Common and distinctive features in stimulus similarity: A modified version of the contrast model. *Psychonomic Bulletin and Review*, 11, 961-974.

Nilson, H.H. Olsson, H. & Juslin, P. (2005). The Cognitive Substrate of Subjective Probability. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 31 (4), pp. 600-620.

Nisbett, R. & Ross, L. (1980). *Human Inference: Strategies and shortcomings of social judgment*. NJ: Prentice-Hall Inc.

Piattelli-Palmarini, M. (1996). *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*. New York: Wiley.

Payne, J.W., Bettman, J. R. & Johnson, E. J. (1993). *The Adaptive Decision Maker*. NY: Cambridge University Press.

Plous, S. (1993). *The psychology of judgment and decision making.* Philadelphia: Temple University Press.

Quine, W. V. (1969). Natural kinds. In W. V. *Quine Ontological Relativity & Other Essays*. New York: Columbia University Press.

Rozoff, D. (1964). Heuristic. *The Accounting Review* 39 (3), 768-769.

Samuels, R., Stich S. & Bishop, M. (2002). Ending the Rationality Wars: How to Make Normative Disputes about Cognitive Illusions Disappear in Elio, R. (ed.) *Common Sense, Reasoning and Rationality*. New York: Oxford University Press.

Shafir E., Smith, E.E. & Osherson, D. (1990). Typicality and reasoning fallacies. *Memory and Cognition* 18 (3): 229-239.

Smith, L.B. (1989). A model of perceptual classification in children and adults. *Psychological Review* 96, 125-144.

Smith, J.D. & Kemler Nelson, D.G. (1984). Overall similarity in adults' classification: The child in all of us. *Journal of Experimental Psychology: General* 113, 137-159.

Sutherland, S. (1992). *Irrationality: The enemy within*. Constable.

Tversky, A. & Gati, I. (1978). Studies of similarity. In E. Rosch & Lloyd, B. (eds.). *Cognition and Categorization*, Hillsdale, NJ: Erlbaum.

Tversky, A. & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Villejoubert, G. & Mandel, D.R. (2002). The inverse fallacy: An account of deviations from Bayes's theorem and the additivity principle. *Memory and Cognition* 30 (2) 171-178.

Ward, T.B. (1983). Response tempo and separable-integral responding: Evidence for an integral-to separable processing sequence in visual perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9, 103-l 12.

White, M. (2006). The Ghost Bird, Ivory-billed Woodpecker. *National Geographic Magazine*, December.

Table 1: Accuracy from using the likelihood heuristic and incremental accuracy from using Bayes' rule

p′: Prior Probability	***H:* Hypothesis set**				
	{.5,.5}	{.6,.4}	{.7,.3}	{.8,.2}	{.9,.1}
[.5,.5]	.00	.00	.00	.00	.00
[.6,.4]	.10	.00	.00	.00	.00
[.7,.3]	.20	.06	.00	.00	.00
[.8,.2]	.30	.13	.04	.00	.00
[.9,.1]	.40	.23	.09	.03	.00
[.95,.05]	.45	.27	.12	.04	.00
$A(\mathbf{L},H,\mathbf{p})$	.50	.68	.84	.94	.99

To obtain Bayesian accuracy for each cell, add the incremental accuracy to $A(\mathbf{L},H,\mathbf{p})$. For instance, when $H=\{.6,.4\}$, and $\mathbf{p}'=[.8,.2]$, the accuracy of the likelihood heuristic is .68 and the accuracy of Bayes' rule is $A(\mathbf{B},H,\mathbf{p})=.68+.13=.81$.

Table 2: Mean inter-subject correlation between similarity judgments, both intra- and inter-context

Set	Similarity	Similarity/ Population	Inter-context	Overall
1	.79	.69	.68	.71
2	.67	.76	.72	.72
3	.85	.73	.79	.79
4	.76	.69	.72	.72

Table 3: Correlations between individual choices by "quasi-subjects" in the four conditions (N=1200). P<.001 except *p<.01.

		Similarity/ Population	Choice/No Prior	Choice/ Prior
First-order correlations	**Similarity**	0.67	0.66	0.61
	Similarity/Population	--	0.61	0.59
	Choice/No Prior		--	0.61
LKChoice, PrChoice and BayesChoice partialled out	**Similarity**	0.26	0.21	0.11
	Similarity/Population	--	0.12	*0.07
	Choice/No Prior		--	0.12

Table 4: Correlations between median RTs in the four conditions

	Similarity/ Population	Choice/No Prior	Choice/ Prior
Similarity	0.66	0.51	0.68
Similarity/Population	--	0.63	0.76
Choice/No Prior		--	0.74

Table 5: A cross-tabulation between choices based on the average similarity judgment and the majority choice for triples, in Bar Hillel's 1974 study and in ours

Bar-Hillel (1974)			
		Choice	
		Pop L	Pop R
Similarity	Pop L	11	0
	Pop R	4	13
φ = .75			

Our data			
		Choice	
		Pop 1	Pop 2
Similarity	Pop 1	54	3
	Pop 2	3	60
φ = .90			

Figure captions

Figure 1: Typical stimuli used by Bar-Hillel (1974). The dashed line in Panel L is not in the original.

Figure 2: Stimuli consisting of two populations of 100 rectangles and a sample of 25 rectangles.

Figure 3: The proportion of times that Population 2 would be chosen by Bayes' rule, as a function of the 9-point similarity scale.

Figure 4: The proportion of correct choice predictions for each respondent in the two choice groups.

Figure 5: Boxplots of median RT in the four conditions.

Figure 6: Median response time plotted against average Similarity judgment for both choice conditions.

Figure 7: Accuracy (BayesChoice) as a function of consistency between prior probability and correct choice.

Figure 8: Proportional shift statistic (PSS) as a function of the mean similarity rating for individual questions.

Figure 9: Lens model adapted from Brunswik.

FIG 1

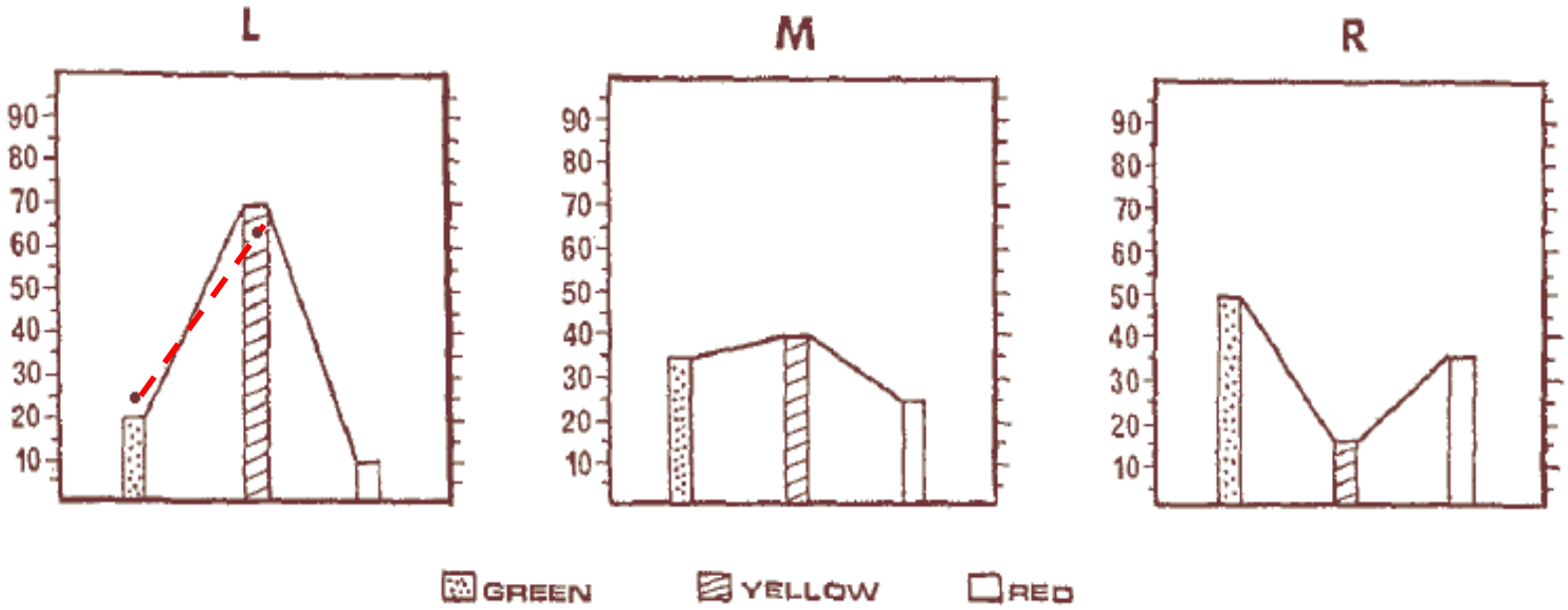
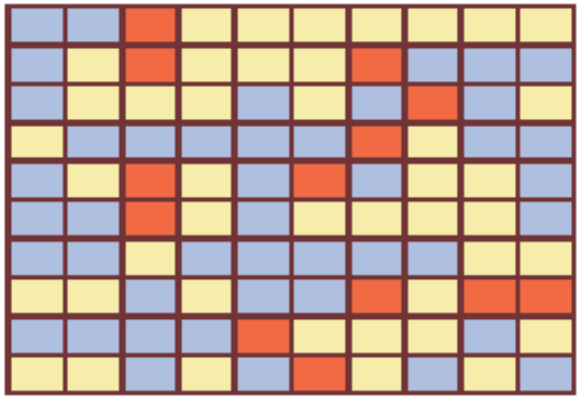
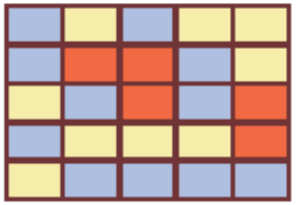


FIG 2





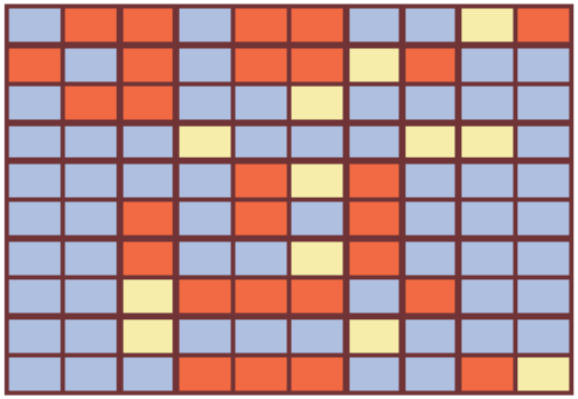


FIG 3

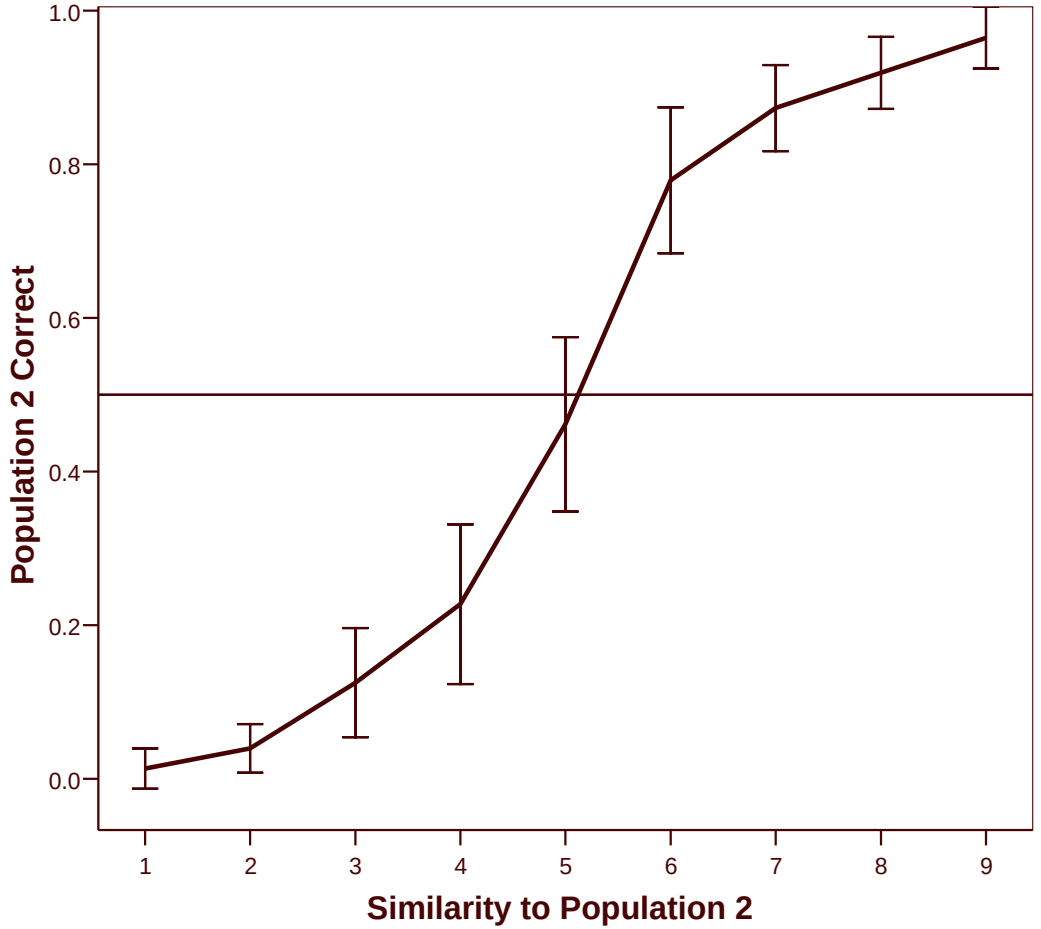


FIG 4

Choice/No prior **Choice/Prior**

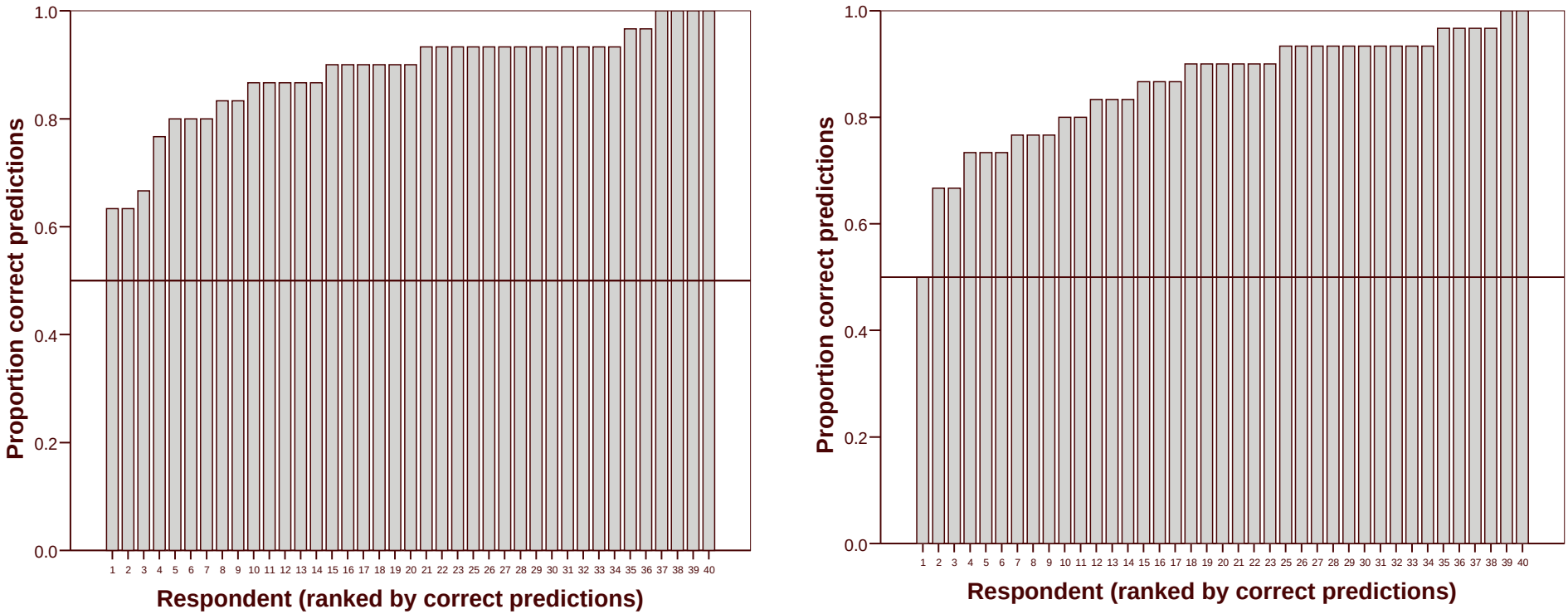

Proportion correct predictions
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Respondent (ranked by correct predictions)
Proportion correct predictions
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
Respondent (ranked by correct predictions)

FIG 5

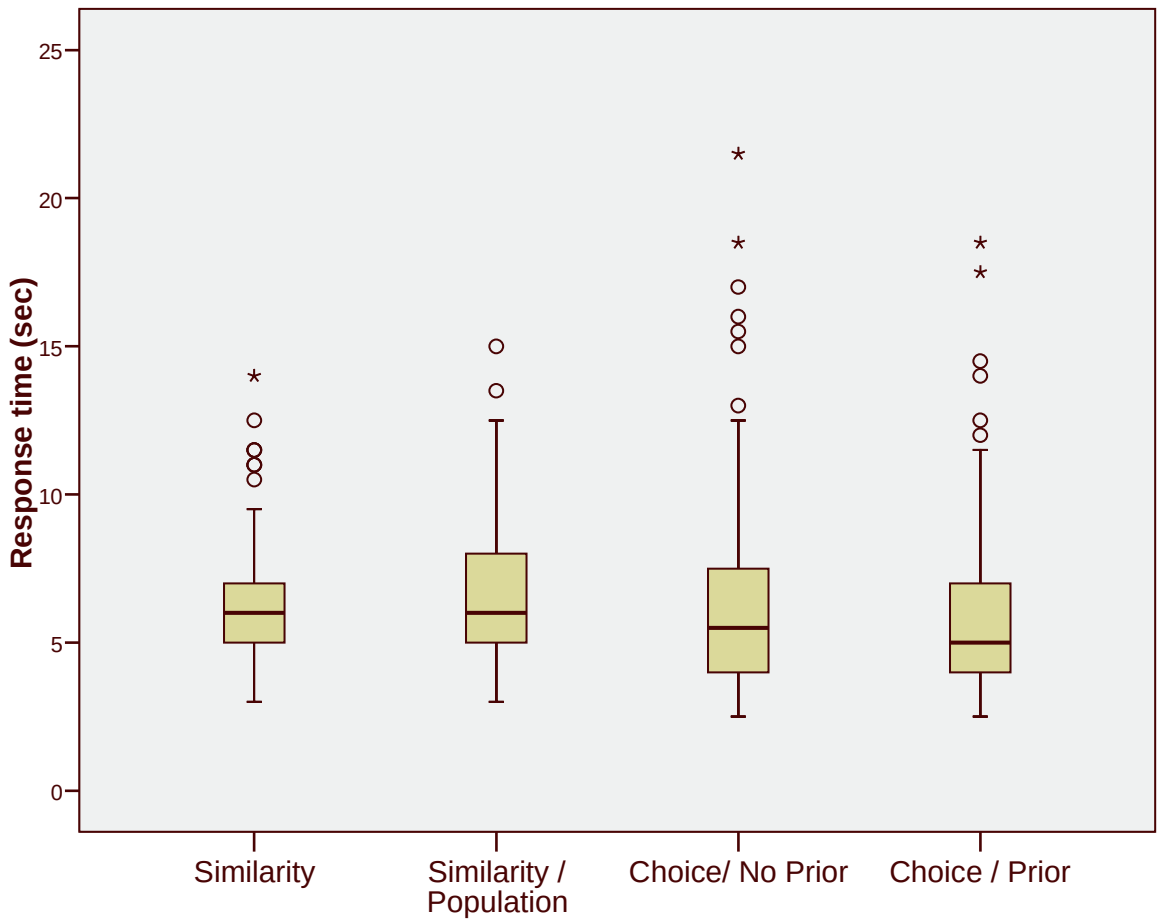
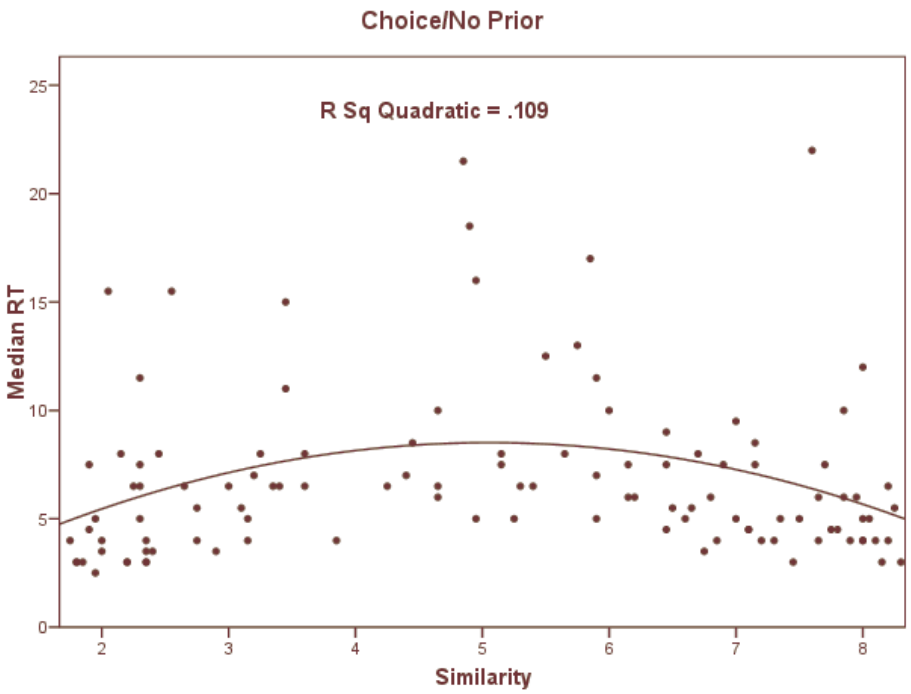

Response time (sec)
25
20
15
10
5
0
Similarity
Similarity / Population
Choice/ No Prior
Choice / Prior

FIG 6


Choice/No Prior
R Sq Quadratic = .109
Median RT
Similarity
0
5
10
15
20
25
2
3
4
5
6
7
8

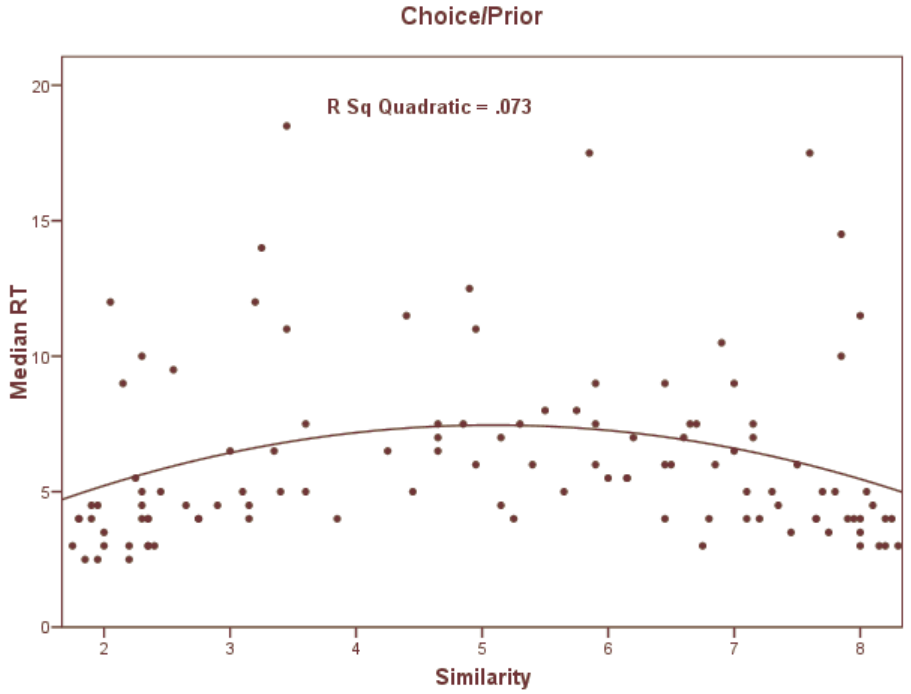

Choice/Prior
R Sq Quadratic = .073
Median RT
Similarity
0
5
10
15
20
2
3
4
5
6
7
8

FIG 7

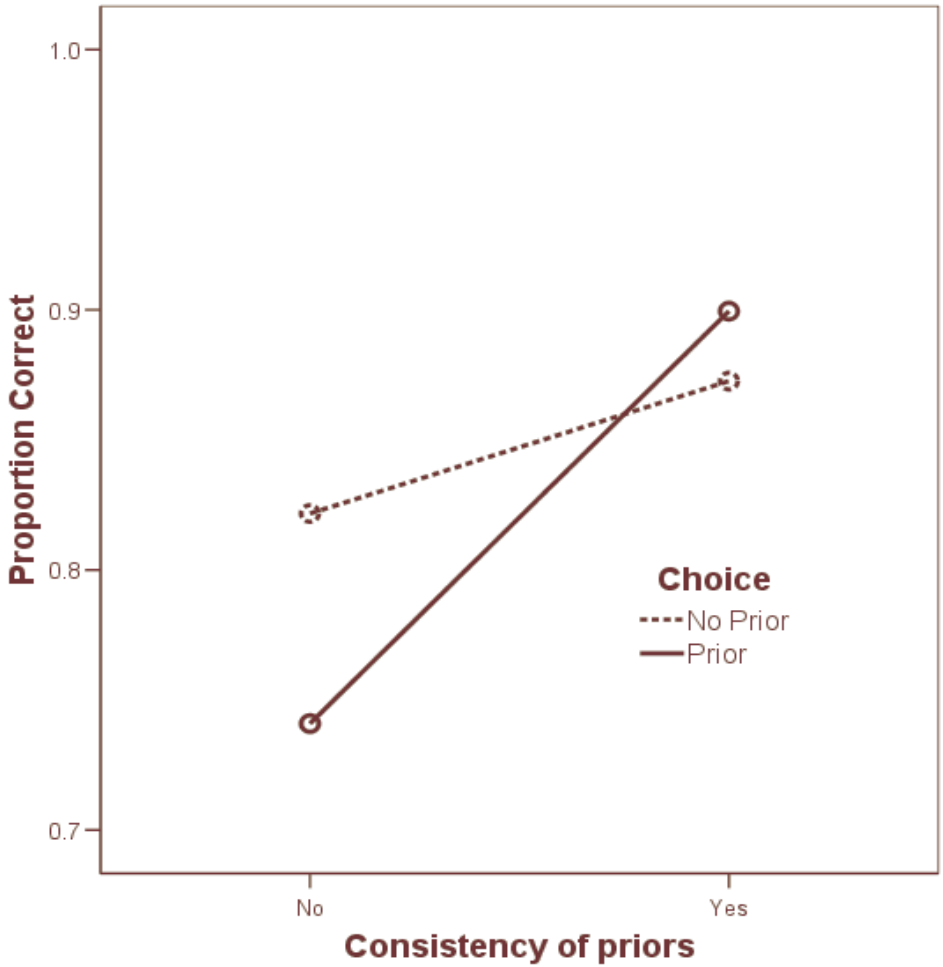

1.0
0.9
0.8
0.7
Proportion Correct
Choice
No Prior
Prior
No
Yes
Consistency of priors

FIG 8

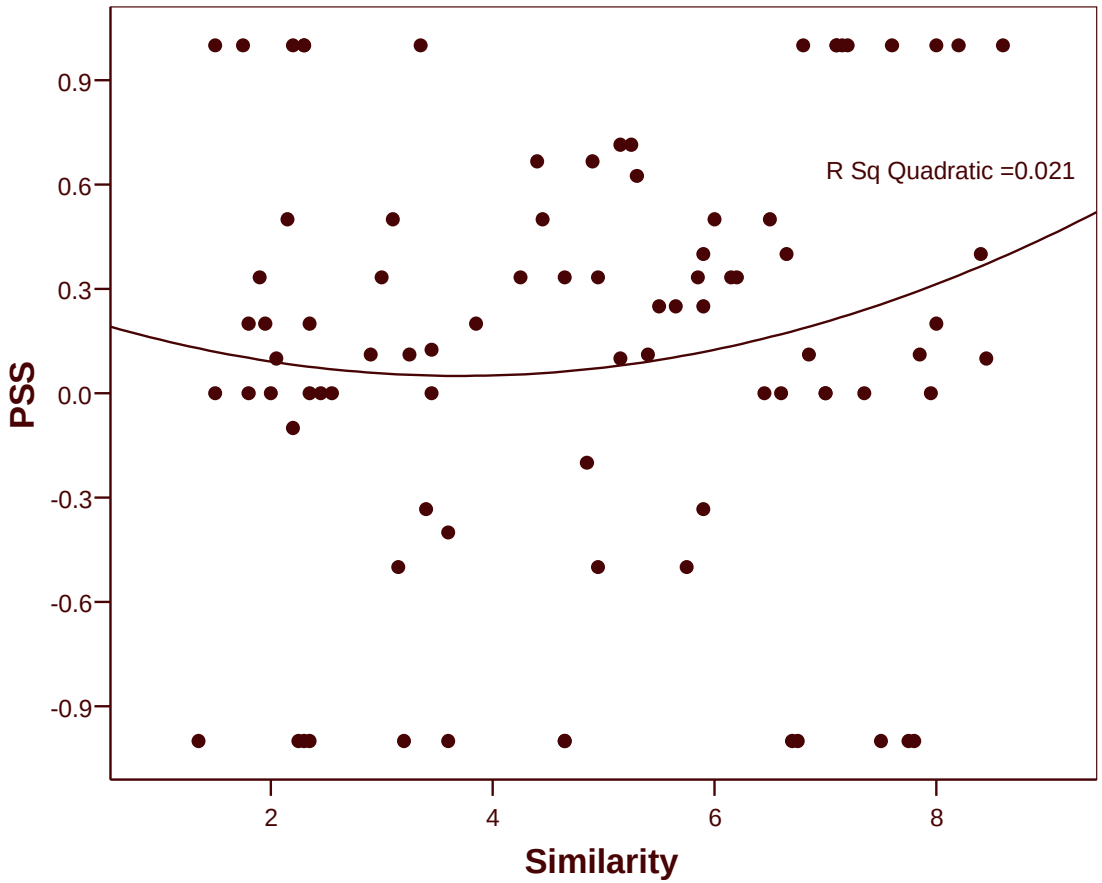
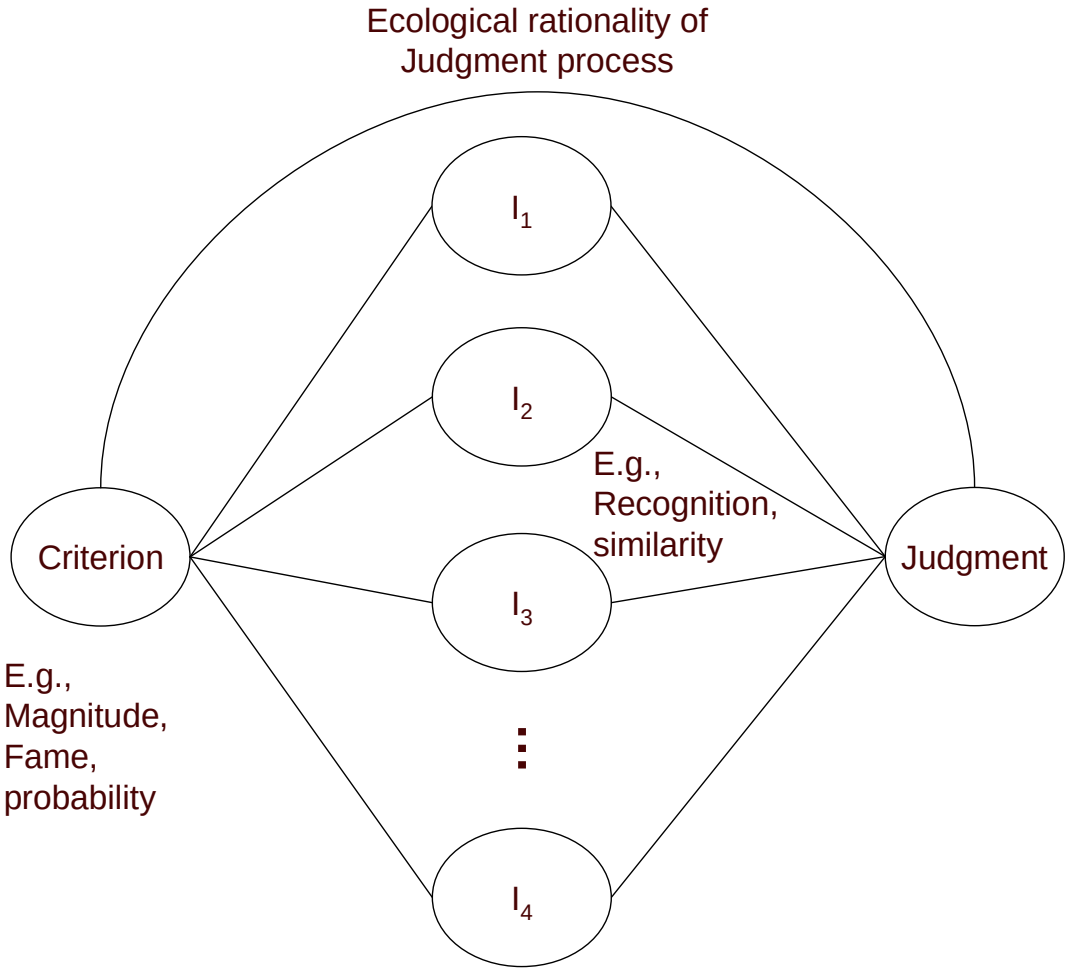

R Sq Quadratic =0.021
PSS
0.9
0.6
0.3
0.0
-0.3
-0.6
-0.9
2
4
6
8
Similarity

FIG 9


Ecological rationality of
Judgment process
I_1
I_2
I_3
I_4
E.g.,
Recognition,
similarity
Criterion
Judgment
E.g.,
Magnitude,
Fame,
probability

Endnotes

[i] This is a further demonstration of the availability heuristic in action. If the only probability judgments we can remember are the 'Linda' or 'Taxicab' problem, then we might well overestimate the frequency with which such erroneous judgments are made.

[ii] Gilovich & Griffin (2003, p.8) observe that 'studies in this [heuristics and biases] tradition have paid scant attention to assessing the overall ecological validity of heuristic processes…assessing the ecological validity of the representativeness heuristic would involve identifying a universe of relevant objects and then correlating the outcome value for each object with the value of the cue variable for each object… . This Herculean task has not attracted researchers in the heuristics and biases tradition; the focus has been on identifying the cues that people use, not on evaluating the overall value of those cues.'

[iii] The term has been used before. Medin, Goldstone and Gentner (1993) use it to refer to the use of similarity as a guide to making 'educated guesses' in the face of uncertainty, a view which closely reflects our own. Kahneman and Frederick (2002) used the term as an alternative label for the representativeness heuristic itself.

[iv] In a simulation study, we found only 0.3% of possible stimuli have all four properties of Bar-Hillel's samples.

[v] Similarity is a complex judgment and in this paper we do not consider *how* it is assessed. For recent candidate models of similarity judgment see Kemp, Bernstein and Tenenbaum, 2005, and Navarro and Lee, 2004.

[vi] The damping parameter adopted by Nilsson et al. (2005; see their Eq. (2)) can be incorporated by introducing a further stage in the model, between the similarity vector and maximum similarity vector.

[vii] Condition (c) is always applicable to our analysis, since the prior probability of all hypotheses other than Urn *A* or Urn *B* is 0.

[viii] This analysis cannot be interpreted as showing how much the similarity heuristic is contributing to choice. Rather, similarity judgments work *because* they are highly correlated with the statistical basis for choice and therefore when we partial out *LKChoice* and *BayesChoice*, we are also partialling out the factors that make it a good decision rule. The analysis is rather a decisive demonstration that we cannot say respondents are "merely" computing Bayesian posterior probabilities and responding accordingly.

[ix] The linear function accounted for none of the variance in median RT, and a cubic function yielded identical fit to the quadratic.

[x] This is a general result. If there are *n* hypotheses to be tested, the similarity heuristic calls on *2n-1* EIPs (*n* calculations and *n-1* comparisons), while the normative rule calls on *4n-1* EIPs (*2n* calculations, *n* products, and *n-1* comparisons).

[xi] Much of the debate revolves around a fuzzy film in which a woodpecker is seen in the distance for 4 seconds (e.g. Fitzpatrick et al., 2005). Given the extremely low prior probability that any ivory-billed woodpecker is alive, it could be argued that even under its best interpretation this evidence could *never* warrant concluding that the posterior probability is appreciably greater than zero.