12

Assessing Evidential Support in Uncertain Environments

Chris M. White and Derek J. Koehler

In this chapter, we explore a classic problem in psychology: How do individuals draw on their previous experience in an uncertain environment to make a prediction or diagnosis on the basis of a set of informational cues? Intuitive predictions based on multiple cues are often highly sensitive to environmental contingencies yet at the same time often exhibit pronounced and consistent biases.

Our research has focused on judgments of the probability of an outcome based on binary cues (e.g., present/absent), in which the diagnostic value of those cues has been learned from direct experience in an uncertain environment. For example, in medical diagnosis, we might predict which disease a patient has based on a symptom that a patient does or does not exhibit. We have developed a model that captures both the strengths and weaknesses of intuitive judgments of this kind (the Evidential Support Accumulation Model, ESAM; Koehler, White, & Grondin, 2003). The model assumes that the frequency with which each cue is observed to have co-occurred with the outcome variable of interest is stored. The perceived diagnostic value of each cue, based on these frequencies, is calculated in a normatively appropriate fashion and integrated with the prior probability of the outcome to arrive at a final probability judgment for a given outcome.

ESAM is designed to account for people's behavior in uncertain environments created in the lab that are, effectively by definition, sampled without bias. In addition, we believe that these types of uncertain environments (those in which multiple cues imperfectly predict which of multiple outcomes will occur) can be viewed as a representative sample of the environments that people actually encounter in their lives. (In addition, we sample a number of different environments in the simulations that we report.) Therefore, factors of external sampling cannot explain the observed judgmental biases. The most important of these biases is that probabilities assigned to competing hypotheses given a particular pattern of symptoms

consistently sum to more than 1 and that this sum increases as the number of present symptoms increases.

Since external sampling cannot explain these biases, we attribute them to the effects of internal sampling. In this chapter we show that when the amount of information that the model samples is reduced, relatively small decrements in performance occur. We also summarize previous research showing that sampling biases incorporated into the model allow it to reproduce a number of effects exhibited in people's judgments. Therefore, based on the current results, we conclude that the internal sampling biases posited by the model may not be overly harmful to the accuracy of the judgments it produces in at least some environments and judgment tasks.

The model is selective with regard to the sampling of informational cues in three respects. First, ESAM disregards conditional dependencies among cues. Second, the model is selectively attuned to present rather than absent cue statuses. And third, a more restricted set of cues is sampled when assessing the extent to which the evidence supports the alternative hypotheses than when assessing the support for the focal hypothesis of the judgment. As a result, judgments produced by the model are systematically biased even when they are based on experience in the form of an objectively representative information sample.

In summary, our work concentrates on the selective internal sampling of cue information. The unit of sampling is at the cue level, which is the lowest level at which it could realistically be in such a task, as is the case for the "Take the Best" heuristic (Gigerenzer and Goldstein, 1999). Toward the end of this chapter we discuss work by other researchers who assume that the unit of internal sampling for this type of task is that of entire exemplars (Dougherty, Gettys, & Ogden, 1999; Juslin & Persson, 2002).

This chapter is organized as follows: After describing ESAM and showing how certain cues are selected for sampling, we review recent experimental results that are consistent with the basic predictions of the model. Following this, we explore how ESAM's performance varies (relative to that of a Bayesian model) based on the number of cues that are sampled in different types of multiple-cue learning environments. We then briefly review the work of other researchers in this area and offer some general conclusions.

THE MODEL

Support Theory

ESAM was developed within the theoretical framework of support theory (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997). According to support theory, judged probability is assigned to descriptions of events, referred to as *hypotheses*, rather than directly to set-theoretic events as in

probability theory. Support theory is thus nonextensional, allowing different probability judgments to be assigned to different descriptions of the same event, reflecting the observation that intuitive judgments of an event's probability are systematically influenced by the way in which that event is described (e.g., Fischhoff, Slovic, & Lichtenstein, 1978).

Support theory consists of two basic assumptions. The first is that judged probability reflects the relative support for the focal and alternative hypotheses:

$$P(A, B) = \frac{s(A)}{s(A) + s(B)}. \tag{12.1}$$

That is, the judged probability of focal hypothesis A rather than alternative hypothesis B is given by the evidential support for A, denoted $s(A)$, normalized relative to that available for B, denoted $s(B)$. If, for example, A and B represent two mutually exclusive diseases from which a patient might be suffering, the judged probability that the patient has disease A rather than disease B, denoted $P(A, B)$, is assumed to reflect the balance of evidential support for A versus that for B.

Support theory's second assumption is that if A is an implicit disjunction (e.g., the patient has a respiratory infection) that refers to the same event as an explicit disjunction of exclusive hypotheses A_1 and A_2 (e.g., the patient has a viral respiratory infection or a bacterial respiratory infection), denoted $(A_1 \vee A_2)$, then

$$s(A) \leq s(A_1 \vee A_2) \leq s(A_1) + s(A_2). \tag{12.2}$$

That is, the support of the implicit disjunction A is less than or equal to that of the explicit disjunction $A_1 \vee A_2$ (because how an event is described affects its assessed support), which in turn is less than or equal to the sum of support of its components when assessed individually (Rottenstreich & Tversky, 1997). In short, unpacking the implicit disjunction A into its components A_1 and A_2 can only increase its support, and hence its judged probability (cf. Fischhoff et al., 1978). The relationship between the support of A and its components A_1 and A_2 is said to be *subadditive*, in the sense that the whole receives less than the sum of its parts.

Support theory implies that, whenever an elementary hypothesis is evaluated relative to all of its alternatives taken as a group (referred to as the *residual*), the weight given to an alternative included implicitly in the residual is generally less than what it would have received had it been evaluated in isolation. Consider a case in which there are three elementary hypotheses: A, B, and C. For instance, suppose a patient is suffering from one (and only one) of three possible flu strains. According to support theory, when a person is asked to judge the probability that the patient is suffering from Flu Strain A, the resulting "elementary" probability judgment $P(A, \overline{A})$ is determined by the evidential support for Flu Strain A normalized

relative to that for its complement (the residual not-A, represented $\overline{A}$). In this case, its complement is an implicit disjunction of Flu Strains B and C. Support theory implies that packing these alternatives together in an implicit disjunction (i.e., the residual) generally results in a loss of support, thereby increasing A's judged probability.

As a result, if separate elementary judgments are obtained of the probability of hypotheses A, B, and C, the total probability

$$T = P(A, \overline{A}) + P(B, \overline{B}) + P(C, \overline{C}) \tag{12.3}$$

assigned to the three elementary hypotheses will generally exceed 1, in violation of probability theory. The degree of subadditivity associated with the set of elementary judgments can be measured by the extent to which the total probability T assigned to them exceeds 1; the greater the value of T, the greater the degree of subadditivity.

A more precise measure of the degree of subadditivity associated with a single judgment is given by a discounting factor $w_{\overline{A}}$ that reflects the degree to which support is lost by packing individual hypotheses into the residual $\overline{A}$:

$$s(\overline{A}) = w_{\overline{A}}[s(B) + s(C)]. \tag{12.4}$$

Support theory's assumption of subadditivity (12.2) implies $w_{\overline{A}} \leq 1$. Lower values of $w_{\overline{A}}$ reflect greater subadditivity, that is, greater loss of support as a result of packing hypotheses B and C into the residual $\overline{A}$.

Koehler, Brenner, and Tversky (1997) offered a simple linear-discounting model according to which the support for the alternatives included in the residual is discounted more heavily as the support for the focal hypothesis of the elementary judgment increases; in other words, $w_{\overline{A}}$ decreases as $s(A)$ increases. This model captures the intuition that when the focal hypothesis is well supported by the available evidence, people feel less compelled to consider how the evidence might also support its alternatives than when the focal hypothesis is not well supported by the evidence. We refer to this phenomenon as *enhanced residual discounting*.

ESAM

Support theory describes the translation of support into probability but does not specify how support is assessed in the evaluation of the available evidence, which is assumed to vary across different judgment tasks and domains. Recently, we developed ESAM as a model of the support assessment process underlying judgments of probability based on patterns of binary (present/absent) cues, where the diagnostic value of those cues and the base rate of the relevant outcomes have been learned from previous experience in an uncertain environment (Koehler et al., 2003).

The model can be thought of as characterizing how a sample of information (i.e., previous experience) obtained in an uncertain environment is represented and subsequently used to assess the likelihood of a particular outcome. More specifically, it is assumed that previous experience in the environment is represented in the form of frequency counts of co-occurrences between informational cue values and outcomes of interest to the judge. Hence a sample of information, in these terms, consists of a set of counts of how frequently a particular outcome has been observed in conjunction with each available informational cue value. For simplicity, we assume that these frequency counts, obtained from experience in the probabilistic cue-based environment, are encoded and later retrieved without error (although this assumption could obviously be refined based on research on frequency estimation; see Sedlmeier & Betsch, 2002).

Because ESAM evaluates the available evidence on a cue-by-cue basis rather than in terms of the entire cue pattern as a whole, it takes as input the co-occurrence frequency with which each cue value (present or absent) has been observed in the presence of each possible hypothesis (or outcome) under evaluation. Although ESAM can readily accommodate any number of hypotheses and cues (and cues with more than two possible values), we will focus on the case of three possible hypotheses and six binary cues. This case corresponds to several experiments, reviewed in the "Experimental Tests" section, involving the diagnosis of simulated "patients" suffering from one of three possible flu strains on the basis of the presence or absence of six discrete symptoms. In this case, the model requires for each of the six cues a count of how frequently that cue was observed as being present and as being absent with each of the three possible hypotheses (i.e., how often a particular symptom was present or absent in conjunction with each of the three possible flu strains). Here, a symptom's absence refers to when a patient is known to not have the symptom; it does not refer to the situation in which the status of the symptom is unknown.

The frequency with which cue C is observed as present in cases where hypothesis H holds is denoted $f_1(C, H)$; the frequency with which cue C is observed to be absent in cases where hypothesis H holds is denoted $f_0(C, H)$. The number of competing hypotheses (or possible outcomes) is denoted N_H and the number of available cues is denoted N_C. Finally, let $f(H)$ represent the overall frequency, or "base rate," with which hypothesis H holds in the set of stored observations.

Diagnostic Value of Each Cue. ESAM assumes that the cue pattern serving as the basis of the probability judgment is assessed one cue at a time, with the diagnostic implication of each observed cue value being evaluated with respect to a target hypothesis. In accord with the Bayesian approach to subjective probability, a piece of evidence is said to be diagnostic with respect to a hypothesis to the extent that the introduction of the evidence

justifies a change in the probability of that hypothesis relative to its prior or base-rate probability.

The diagnostic value $d_1(C, H)$ of the presence of cue C with respect to a particular hypothesis H is given as follows:

$$d_1(C, H) = \frac{f_1(C, H)/f(H)}{\sum_j \left[f_1(C, H_j)/f(H_j)\right]}. \tag{12.5}$$

The diagnostic value $d_0(C, H)$ of the absence of cue C with respect to a particular hypothesis H is given by a parallel expression (and if there were more than two cue statuses, more parallel expressions could be used for each of the statuses). The value of d varies between 0 and 1. This calculation can be thought of as distributing one "unit" of diagnostic value among the set of competing hypotheses, with hypotheses implicated by the cue's presence receiving a larger share than hypotheses upon which the cue's presence casts doubt.

In this manner, the model assumes that the judge is sensitive to the diagnostic value of individual cues without necessarily being sensitive to the diagnostic value of cue patterns. Specifically, this calculation of diagnostic value is insensitive to conditional dependence among cues. Consequently, the model will not capture configural cue processing effects (Edgell, 1978, 1980). The calculation of diagnostic value is also uninfluenced by the base rate or prior probability of the hypothesis in question, as is the case for the likelihood ratio in Bayes's rule.

According to ESAM, the diagnostic implication of each cue value constituting the cue pattern is individually assessed and then summed to arrive at an overall assessment of the diagnostic value of the cue pattern taken as a whole for a particular hypothesis. It is assumed that, because of their higher salience, present cue values are given greater weight than are absent cue values in the summation process. The diagnostic value of cue pattern **C** for hypothesis H, denoted $d_{\mathbf{C}}(H)$, is given by

$$d_{\mathbf{C}}(H) = (1-\delta) \sum_i^{\text{present cues}} d_1(C_i, H) + \delta \sum_i^{\text{absent cues}} d_0(C_i, H) \quad \text{for cues } C_i,\ i = 1, \ldots, N_C \text{ in } \mathbf{C}. \tag{12.6}$$

The free parameter δ (which can vary between 0 and 1) represents the weight placed on absent cues relative to that placed on present cues in the cue pattern. Relative underweighting of absent cues is indicated by $\delta < 1/2$ with $\delta = 0$ representing the special case of complete neglect of absent cues. As well as δ being able to represent the amount of weight placed on each set of cues, it could also represent the relative probabilities of cues of each status being sampled when assessing support for a hypothesis. We interpret

this parameter in the latter sense in the Simulations section of this chapter. Note that the value of $d_{\mathbf{C}}(H)$ will generally tend to increase with the number of cues constituting the cue pattern, with a maximum value of N_C.

In general, the weight placed on cues of each status when assessing support for a hypothesis depends on the salience of each status. Therefore, when cues have substitutive statuses (e.g., gender) they will likely have equal salience and therefore receive equal weighting when assessing support. In addition, when cues have more than two statuses, one could either estimate more weighting factors or simply assume equal weighting.

Base-Rate Sensitivity. ESAM's diagnostic value calculation is insensitive to the base rates of the competing hypotheses because it controls for the base-rate frequency $f(H)$ of the hypothesis under evaluation. ESAM accommodates potential base-rate sensitivity of the support assessment process in the translation from diagnostic value to support. The support for hypothesis H conveyed by cue pattern $\mathbf{C}$, denoted $s_{\mathbf{C}}(H) \geq 0$, is given by

$$s_{\mathbf{C}}(H) = \left[\alpha \left(\frac{f(H)}{\sum_j f(H_j)} - \frac{1}{N_H} \right) + (1 - \alpha) d_{\mathbf{C}}(H) \right]^{\gamma}. \tag{12.7}$$

The free parameter α (which can vary between 0 and 1) provides a measure of the extent to which the support for hypothesis H, which is determined primarily by the diagnostic value $d_{\mathbf{C}}(H)$ of the cue pattern for that hypothesis, is adjusted in light of its base rate (i.e., observed relative frequency in comparison with the alternative hypotheses). The adjustment is additive, as Novemsky and Kronzon (1999) found to be the case in human judgments, rather than multiplicative, as it is in the Bayesian calculations. The adjustment is positive in the case of high-base-rate hypotheses whose relative frequency exceeds $1/N_H$, the value expected under a uniform partition; the adjustment is negative for low-base-rate hypotheses. This convention was implemented for computational simplicity because in the special case of equal base rates this adjustment is zero and the parameter α drops out of the model. With unequal base rates, α reflects the judge's sensitivity to this consideration.

After combining the diagnostic value $d_{\mathbf{C}}(H)$ of the cue pattern $\mathbf{C}$ in implicating hypothesis H with an adjustment in light of H's base rate, the resulting value is then exponentiated to arrive at the support for the hypothesis conveyed by the cue pattern (cf. Tversky & Koehler, 1994). The exponent γ is a free parameter (which can take any positive value) that influences the extremity of the resulting judgments. Its value can be interpreted as a measure of judgmental confidence, that is, the confidence with which the judge relies on his or her previous experience in evaluating the evidence.

Assessing Support of the Residual Hypothesis. Recall that, according to support theory, the residual hypothesis (i.e., the collection of alternatives to the focal hypothesis) receives less support than the sum of the support its component hypotheses would have received had they been evaluated individually. Koehler et al. (1997) suggested that the judge is less likely to fully evaluate the extent to which the evidence supports alternative hypotheses as support for the focal hypothesis increases. In ESAM, this assumption of enhanced residual discounting is implemented by restricting the number of cues that are sampled in accumulating support for (alternatives included in) the residual. Specifically, in contrast to the computation of support for the focal hypothesis, in which the diagnostic value of each cue in the cue pattern makes a contribution, it is assumed that only a subset of cues are sampled with regard to their contribution to the support for an alternative hypothesis included in the residual.

For simplicity, we assume that, given a particular level of support for the focal hypothesis, the probability q that a cue will be sampled and its diagnostic value added in the calculation of support for each alternative hypothesis included in the residual is the same for all of the cues. Since we fit the model to judgments aggregated across multiple trials and/or participants, this probability can be implemented in the form of a discounting weight that reflects the proportion of its full diagnostic value, on average, that a given cue will contribute to the support for a hypothesis included in the residual. This discounting weight, denoted $q_{\overline{H}}$ is assumed to be inversely proportional to the support for the focal hypothesis H:

$$q_{\overline{H}} = \frac{1}{\beta\, s(H) + 1}. \tag{12.8}$$

The free parameter β (which can take any nonnegative value) determines how quickly $q_{\overline{H}}$ decreases as $s(H)$ increases.

Support for the residual is then given by

$$s_{\mathbf{C}}(\overline{H}) = \sum_{H_j \text{ in } \overline{H}} \left[\alpha \left(\frac{f(H_j)}{\sum_i f(H_i)} - \frac{1}{N_H} \right) + (1-\alpha) q_{\overline{H}} d_{\mathbf{C}}(H_j) \right]^{\gamma}. \tag{12.9}$$

In other words, the support for each alternative hypothesis included in the residual is determined by the sum of the diagnostic value contributed by each cue, which is discounted to reflect the restricted set of cues consulted in evaluating support for hypotheses included in the residual. The support for the residual as a whole is given by the sum of the support thus calculated of each alternative hypothesis that it includes.

The discounting can be implemented in this way because we are only interested in aggregated judgments. If we were to attempt to model individual judgments on a trial-by-trial basis, with probability $q_{\overline{H}}$ each cue

would either be sampled or not when assessing the diagnostic implications of the cue pattern for the alternative hypotheses. The free parameter β therefore affects how many cues are sampled when assessing the support for the alternative hypotheses.

In summary, ESAM describes how the support for a hypothesis is assessed on the basis of probabilistic cues when information regarding the usefulness of the cues is available from previous observations represented in the form of stored frequency counts. The model has four free parameters: α (base-rate adjustment), β (enhanced residual discounting), γ (judgmental extremity), and δ (absent cue weighting). According to the model, the diagnostic value of each cue for a particular hypothesis is assessed independently and then summed over the cues constituting the cue pattern. The free parameter δ reflects the relative probability of sampling present versus absent cues. If a cue is sampled, its diagnostic value is included in the summation process. The free parameter δ therefore represents the differing psychological saliencies of each cue status. The free parameter α reflects the extent to which the diagnostic value assessment is adjusted in light of the base rate of the hypothesis under evaluation. The free parameter γ reflects the extremity of the resulting support estimates and reflects the confidence that a person has in his or her assessment of the evidence. The free parameter β reflects the degree to which, as the support for the focal hypothesis increases, the set of cues sampled in assessing the support for hypotheses included in the residual is restricted. We can therefore see that the free parameters δ and β control what proportion of the available information is sampled when computing the probability for the focal hypothesis given a certain pattern of cues. In the Simulations section we find that reasonably accurate responses are generated even when the proportion of information sampled is relatively low.

Normative Benchmark

In what ways would a Bayesian approach differ from that of ESAM in evaluating the implications of a pattern of cues for a particular hypothesis in light of previous experience with those cues? Although there are a number of more or less complicated approaches that could be developed from a Bayesian perspective (e.g., see Martignon & Laskey, 1999), we will consider only the simplest one here, which relies heavily on the assumption of conditional independence of cue values. If the cue values constituting the cue pattern are conditionally independent, then one can readily calculate the probability of observing any particular cue pattern given that a designated hypothesis holds (e.g., the probability of observing a particular pattern of symptoms given that the patient has a designated flu strain) as the product of the conditional probabilities of each individual cue given that hypothesis. This calculation serves as the basis for evaluating the

likelihood ratio in the Bayesian approach, which is then combined with the prior probability of the hypothesis in question to arrive at an assessment of its posterior probability (i.e., its probability in light of the cue pattern). Assuming that both the conditional probabilities of each cue value and the overall prior probability of each hypothesis are estimated from a set of stored frequency counts summarizing previous experience with the cues, the probability of a hypothesis H given a pattern of cue values $\mathbf{C}$ is given by,

$$P(H \mid \text{cue pattern C}) = \frac{f(H)\prod_{i=1}^{N_C}\frac{f(C_i,H)}{f(H)}}{\sum_{j=1}^{N_H}\left[f(H_j)\prod_{i=1}^{N_C}\frac{f(C_i,H_j)}{f(H_j)}\right]} \text{ for } C_i \text{ in cue pattern C,} \tag{12.10}$$

where $f(C_i, H)$ is the frequency with which cue value C_i (absent or present) was previously observed in conjunction with hypothesis H.

The numerator of (12.10) can be viewed as corresponding to the extent to which, in the Bayesian analysis, hypothesis H is supported in light of the available evidence. The product term in the numerator corresponds to the diagnostic value of the evidence, which is adjusted in light of the base rate or prior probability of the hypothesis in question as reflected by $f(H)$. The same calculation is used to assess the support conveyed by the cue pattern for each of the competing hypotheses. As in ESAM, the probability assigned to the hypothesis is given by its normalized support relative to its alternatives. Unlike in ESAM, of course, there is no accommodation in the Bayesian framework for discounting of support arising from packing together the alternatives to the focal hypothesis in the residual. That is, in contrast to ESAM in particular and support theory in general, the normative Bayesian framework produces judgments that are necessarily extensional (i.e., bound by rules of set inclusion) and additive (i.e., over decomposition of events into subsets).

Another key difference between ESAM and the Bayesian model (12.10) just described is that the Bayesian model integrates individual cue values (and considerations of hypothesis base rate or prior probability) in a multiplicative manner, whereas ESAM uses an additive integration form. As a consequence, in ESAM's additive framework, support tends to increase with the number of cues consulted, whereas in the Bayesian model it tends to decrease. Furthermore, in the integration process, ESAM accommodates differential weighting of cue absence and cue presence, whereas in the normative Bayesian approach cue absence and cue presence are logically interchangeable. Because ESAM is not a generalization of the Bayesian model (12.10) outlined here, there are no parameter values for which ESAM will exactly reproduce the corresponding judgments derived

from the Bayesian approach. [A generalization of the Bayesian model is offered by Koehler et al. (2003).] ESAM does tend to produce judgments that correlate highly with the corresponding Bayesian values, however, when $\beta = 0$ and $\delta = {}^1/_2$. The former represents the special case of ESAM that produces additive judgments; and the latter places equal weight – as in the Bayesian model – on cue absence and cue presence. The performance of this version of ESAM, which samples all of the available information, is reported in the first simulation in the Simulations section later in the chapter.

An alternative – also arguably normative – approach adopts a frequentist perspective, in which the current cue pattern serving as evidence is assessed against previous observations that exactly match that pattern. The judged probability of a designated hypothesis is given by the proportion of previous cases matching the cue pattern in which the hypothesis held (e.g., the proportion of previous patients with an identical set of symptoms who suffered from the hypothesized flu strain). This approach represents the starting point in development of exemplar-based models of classification learning (e.g., Brooks, 1978; Medin & Schaffer, 1978) and has the advantage of being able to accommodate cue structures for which conditional dependence does not hold. Both ESAM and the Bayesian model outlined here, by contrast, assume conditional independence of cues. The frequentist approach does, however, require a relatively large sample of previous observations to produce reliable probability estimates. It also requires stored frequency counts for every possible cue pattern, the number of which increases exponentially with the number of cues. As descriptive models, then, either ESAM or the Bayesian model outlined here might be more useful in producing reasonably accurate judgments in the face of small sample sizes and limited memory capacity.

EXPERIMENTAL TESTS

We have found ESAM to closely reproduce probability judgments made in several multiple-cue probability learning experiments conducted by Koehler (2000) and Koehler et al. (2003). All of these experiments employ a simulated medical diagnosis task in which participants attempt to diagnose which of three possible flu strains a patient is suffering from based on the set of symptoms exhibited by that patient.

General Method

These experiments generally proceed in two phases. In an initial learning phase, participants learn about cue–outcome (i.e., symptom–flu strain) associations from direct, trial-by-trial learning experience in the uncertain environment. On a typical learning trial, the participant is presented with a

patient characterized in terms of a set of four to six binary (present/absent) symptoms, and on the basis of this information the participant makes a forced-choice diagnosis of the flu strain from which the patient is suffering. After making their diagnosis, participants are told which flu strain the patient actually has. The learning phase consists of a set of 240–300 such trials, which can be viewed as the information sample drawn from the uncertain environment on which subsequent inferential judgments must be based. Sampling from the uncertain environment is conducted such that participants encounter an unbiased sample presented in a passive, sequential manner. In the second, judgment phase of the experiment, participants are presented with additional patients and, for each, judge the probability that the patient is suffering from a designated flu strain. The judgment phase instructions emphasize that the flu strain designated as the target of judgment on a particular trial is selected arbitrarily and that its designation should not be taken as having any informational value regarding the patient's diagnosis.

Typical Results

Across the set of judgment trials, participants eventually assign a probability to each of the three competing hypotheses (flu strains) contingent on each possible combination of cue values (symptoms). There are two aspects of the results of these studies that are of particular interest. First, as predicted by support theory, the probability judgments tend to be systematically subadditive; that is, the sum of the probabilities assigned to the three competing hypotheses given a particular pattern of cue values consistently exceeds one, in contrast to the requirement of probability theory that the total probability assigned to a mutually exclusive and collectively exhaustive set of events should add to one. Second, the degree of subadditivity for a set of judgments increases with the number of present cue values (symptoms) in the cue pattern on which the judgments are based. That is, the total probability assigned to the three flu strains is higher for patients exhibiting many symptoms than for patients exhibiting fewer symptoms.

Koehler (2000) found these systematic biases to hold even when (a) the training phase required participants to make probability judgments rather than forced-choice diagnoses and (b) the judgment phase required participants to make judgments of absolute frequency rather than of probability or relative frequency. In light of many criticisms of work in the heuristics and biases tradition that judgmental biases arise from nonrepresentative item sampling or use of problems that otherwise contradict the previous experience of research participants, it is notable that we have observed such pronounced and systematic biases in judged probability even when the judgments are based on a directly experienced representative sample of information from the judgment environment.

Fitting ESAM to the Data

ESAM was developed as a model of how probability judgments are generated in this type of task. It could be extended to generate choices or diagnoses, as participants give in the initial learning phase of our studies, and this is done in the Simulations section. However, our previous work only investigated subjective probability judgments.

ESAM readily reproduces the key experimental results from our studies, namely that the probability judgments are generally subadditive and that the degree of observed subadditivity increases with the number of present cue values in the cue pattern on which the judgments are based. That latter result follows from ESAM's assumptions of absent-cue underweighting (i.e., $\delta < {}^1/_2$; absent cue values are given less weight than present cue values) and of enhanced residual discounting (i.e., $\beta > 0$; support for alternative hypotheses included in the residual is discounted more heavily as support for the focal hypothesis increases). The result of these assumptions is that, as the number of present cue values increases, support for the focal hypothesis also tends to increase, and consequently the degree of subadditivity increases (i.e., support for specific alternative hypotheses included in the residual is more heavily discounted).

Koehler et al. (2003) fit ESAM to the judgment data from four multiple-cue probability learning experiments. The best-fitting values of three of ESAM's four free parameters (α was only applicable in one experiment) varied moderately across the four experiments (β varied from 0.10 to 0.66, γ from 1.14 to 2.96, and δ varied least: from 0.22 to 0.27), with some of the variability being accounted for by some of the differences between the learning environments used in each experiment. ESAM was able to account for a large amount of the variability in the observed judgments for which the Bayesian model was unable to account. For example, Figure 12.1 shows the observed versus predicted probability judgments in Experiment 1 (Koehler et al., 2003).

Testing ESAM's Assumptions

Koehler et al. (2003) also compared ESAM's fit to that of a number of alternative models, each of which differed from ESAM by a single key assumption. The alternative models tended to fit the data less well, providing some corroboration for each of ESAM's key assumptions. Specific models exhibiting poorer data fits were those that (a) consider entire cue patterns rather than single cues, (b) fail to place greater weight on cue presence than on cue absence, (c) use different methods for calculating the diagnostic value of a cue, (d) do not adjust for the base rate of a hypothesis, and (e) assume no relationship between the support for the focal hypothesis and discounting of support for its alternatives. Further analyses showed

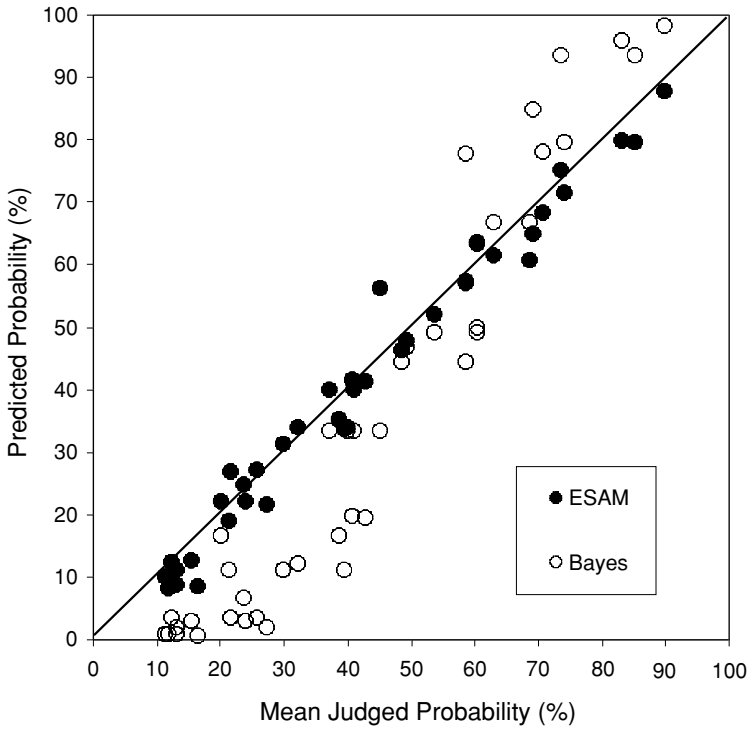

FIGURE 12.1. Mean probability judgments versus ESAM and Bayesian predictions in Experiment 1 of Koehler et al. (2003).

that the model's simplifying assumption of perfectly accurate frequency counts of previous cue–hypothesis co-occurrence proved to cost relatively little in terms of fit to the data.

ESAM's Accuracy

Despite the systematic biases introduced by some of its underlying assumptions, ESAM was found to produce reasonably accurate judgments in the experiments we reported as evaluated by a comparison to the corresponding Bayesian values. The accuracy achieved by ESAM is impressive because the model ignores or underweights some information (e.g., absent cues) used in the Bayesian approach and also uses simpler (e.g., additive instead of multiplicative) rules for combining the implications of different pieces of evidence. The model nonetheless can produce reasonably accurate judgments because it is capable of evaluating and aggregating the diagnostic implications of each cue in the cue pattern. Indeed, the model incorporates the basic definition of diagnosticity used in the Bayesian approach

and, as with Bayes's rule, integrates considerations of the diagnosticity of the evidence and of the base rate or prior probability of the hypothesis. As a result, ESAM and the Bayesian approach tend to produce highly correlated output. This observation is consistent with previous research showing that as long as they correctly identify and integrate the direction of each cue's diagnostic implication, information-integration models will often produce output that corresponds quite closely to the output of the corresponding normative model even if they differ in the weights attached to the cues and in the specific form of the combination rule they employ (e.g., Anderson, 1981; Dawes, 1979).

SIMULATIONS

Following a long line of research in the heuristics and biases tradition, we have developed a descriptive model of subjective probability that uses only a selected subset of the evidence available from the environment as a basis for judgment. As a consequence, its judgments exhibit consistent and predictable biases relative to the standard normative benchmarks. This cost is balanced by the potential benefit of circumventing some of the informational and computational complexities associated with the normative calculations. A natural question one might ask, then, concerns the relative costs and benefits of ESAM's simplifying assumptions. Specifically, we wondered whether it might be possible to characterize the types of judgment environments in which the model renders reasonably accurate judgments (at considerable informational and computational savings) versus the types of environments in which the model performs relatively poorly.

In this section, we first describe the different judgment environments used to assess the relative costs and benefits of ESAM's simplifying assumptions and then define how the performance of ESAM is assessed by comparing it to the output of the Bayesian analysis (12.10). Having established this, we evaluate for each environment how close ESAM's judgments come to the Bayesian benchmark when all of the available information is sampled, and subsequently evaluate the drop in ESAM's performance when only sampling a subset of the information. In general, it is found that ESAM incurs only small costs in performance when not all of the information is sampled, which may be outweighed by the reduction in computational requirements that results.

Environments

We investigated ESAM's performance across a wide range of environments by varying the number of cues, the number of hypotheses, and whether or not all of the cues were equally predictive. We used either

three or six hypotheses and three, six, or twelve conditionally independent cues. In some environments each cue was as predictive as all of the others (we refer to these as the flat cue diagnosticity environments), and in others the predictive value of the cues varied between the cues (we refer to these as the J-shaped cue diagnosticity environments, following Gigerenzer & Goldstein, 1999). The base rates of all of the hypotheses were equal in all environments, and we used binary (present/absent) cues. Experiment 3 by Koehler et al. (2003) used an environment extremely similar to the six-cue, three-hypothesis, J-shaped cue diagnosticity environment used here; the results of that experiment are discussed in the following in light of the current findings.

The diagnostic value of a cue is determined by the proportion of times that the cue is present in conjunction with each of the competing hypotheses, as shown in Tables 12.1 and 12.2 for each of the flat and J-shaped cue diagnosticity environments, respectively. In all environments, a given cue was present (a) most often when one hypothesis occurred (for example, the first cue might be present 85% of the time that hypothesis 1 occurred), (b) the complementary proportion of times when another hypothesis occurred (for example, 15% of the time that hypothesis 3 occurred), and (c) in graded proportions between these two extreme values (for example, between 85% and 15%) when the other hypotheses occurred (for example, if there were only three hypotheses then the cue was present 50% of the time that hypothesis 2 occurred). The frequency with which each of the other cues was present shared this structure, but each cue was present most often in conjunction with a different hypothesis (for example, the second cue might be present 85% of the time that hypothesis 2 occurred, etc.). In this way, a "set" of cues can be constructed with the presence of each cue in the set most strongly implicating a different hypothesis. Therefore, the number of cues in a set equaled the number of hypotheses. Across all of the hypotheses, all cues within a set were equally diagnostic. Only complete sets of cues were used, and so environments in which there would be fewer cues than hypotheses were not possible.

We measured the diagnosticity of each cue by computing the average value of "delta-p" across the hypotheses for that cue (see Wasserman, Dorner, & Kao, 1990). Delta-p is equal to the difference in the conditional probability of a given hypothesis between when the cue is present and when it is absent. We took an average value of this statistic across all of the hypotheses in the environment to obtain a measure of the cue's overall diagnosticity, which we refer to as a "mean delta-p," as defined by the equation,

$$\overline{\Delta P_C} = \left[\sum_i |P(H_i \mid C+) - P(H_i \mid C-)| \right] \Big/ N_H, \tag{12.11}$$

where $C+$ denotes the presence of cue C, and $C-$ its absence.

TABLE 12.1. *Probability of Each Cue Being Present Given the Occurrence of Each Hypothesis in Flat Cue Diagnosticity Environments*

Number of Hypotheses	Number of Cues	Cue	Hypothesis						Mean ΔP
			#1	#2	#3	#4	#5	#6	
3	3	A	0.85	0.50	0.15				0.311
		B	0.15	0.85	0.50				
		C	0.50	0.15	0.85				
3	6	A	0.85	0.50	0.15				0.311
		B	0.15	0.85	0.50				
		C	0.50	0.15	0.85				
		D	0.85	0.50	0.15				0.311
		E	0.15	0.85	0.50				
		F	0.50	0.15	0.85				
3	12	A	0.85	0.50	0.15				0.311
		B	0.15	0.85	0.50				
		C	0.50	0.15	0.85				
		D	0.85	0.50	0.15				0.311
		E	0.15	0.85	0.50				
		F	0.50	0.15	0.85				
		G	0.85	0.50	0.15				0.311
		H	0.15	0.85	0.50				
		I	0.50	0.15	0.85				
		J	0.85	0.50	0.15				0.311
		K	0.15	0.85	0.50				
		L	0.50	0.15	0.85				
6	6	A	0.85	0.71	0.57	0.43	0.29	0.15	0.140
		B	0.15	0.85	0.71	0.57	0.43	0.29	
		C	0.29	0.15	0.85	0.71	0.57	0.43	
		D	0.43	0.29	0.15	0.85	0.71	0.57	
		E	0.57	0.43	0.29	0.15	0.85	0.71	
		F	0.71	0.57	0.43	0.29	0.15	0.85	
6	12	A	0.85	0.71	0.57	0.43	0.29	0.15	0.140
		B	0.15	0.85	0.71	0.57	0.43	0.29	
		C	0.29	0.15	0.85	0.71	0.57	0.43	
		D	0.43	0.29	0.15	0.85	0.71	0.57	
		E	0.57	0.43	0.29	0.15	0.85	0.71	
		F	0.71	0.57	0.43	0.29	0.15	0.85	
		G	0.85	0.71	0.57	0.43	0.29	0.15	0.140
		H	0.15	0.85	0.71	0.57	0.43	0.29	
		I	0.29	0.15	0.85	0.71	0.57	0.43	
		J	0.43	0.29	0.15	0.85	0.71	0.57	
		K	0.57	0.43	0.29	0.15	0.85	0.71	
		L	0.71	0.57	0.43	0.29	0.15	0.85	

TABLE 12.2. *Probability of Each Cue Being Present Given the Occurrence of Each Hypothesis in J-Shaped Cue Diagnosticity Environments*

Number of Hypotheses	Number of Cues	Cue	Hypothesis #1	#2	#3	#4	#5	#6	Mean ΔP
3	6	A	0.85	0.50	0.15				0.311
		B	0.15	0.85	0.50				
		C	0.50	0.15	0.85				
		D	0.675	0.50	0.325				0.155
		E	0.325	0.675	0.50				
		F	0.50	0.325	0.675				
3	12	A	0.85	0.50	0.15				0.311
		B	0.15	0.85	0.50				
		C	0.50	0.15	0.85				
		D	0.675	0.50	0.325				0.155
		E	0.325	0.675	0.50				
		F	0.50	0.325	0.675				
		G	0.5875	0.50	0.4125				0.078
		H	0.4125	0.5875	0.50				
		I	0.50	0.4125	0.5875				
		J	0.544	0.50	0.456				0.039
		K	0.456	0.544	0.50				
		L	0.50	0.456	0.544				
6	12	A	0.85	0.71	0.57	0.43	0.29	0.15	0.140
		B	0.15	0.85	0.71	0.57	0.43	0.29	
		C	0.29	0.15	0.85	0.71	0.57	0.43	
		D	0.43	0.29	0.15	0.85	0.71	0.57	
		E	0.57	0.43	0.29	0.15	0.85	0.71	
		F	0.71	0.57	0.43	0.29	0.15	0.85	
		G	0.675	0.605	0.535	0.465	0.395	0.325	0.070
		H	0.325	0.675	0.605	0.535	0.465	0.395	
		I	0.395	0.325	0.675	0.605	0.535	0.465	
		J	0.465	0.395	0.325	0.675	0.605	0.535	
		K	0.535	0.465	0.395	0.325	0.675	0.605	
		L	0.605	0.535	0.465	0.395	0.325	0.675	

This metric shows how the flat and J-shaped cue diagnosticity environments differed. In flat cue diagnosticity environments, the cues in one set were as diagnostic as the cues in the other sets. In the J-shaped cue diagnosticity environments, the cues in the first set were the most diagnostic and the diagnosticity of the cues in the other sets fell off with a J-shaped distribution. Obviously, J-shaped environments were only possible when there

was at least two sets of cues. It has been argued that most natural environments exhibit this J-shaped pattern of cue diagnosticities (Gigerenzer & Goldstein, 1999), and it is in this type of environment that ignoring many of the cues may not harm the accuracy of the probability judgments to a very large degree if the least diagnostic cues are ignored (as shown by the "Take the Best" heuristic and other frugal algorithms; see Gigerenzer & Goldstein, 1999).

In some of what we call J-shaped cue diagnosticity environments there are only two values of cue diagnosticity. Technically, of course, with only two diagnosticity levels it is not really appropriate to refer to the distribution as J-shaped, as at least three levels would be needed to produce this or any other kind of nonlinear pattern. Therefore, the reader may wish to replace the term J-shaped distribution in these cases with the less specific but more appropriate term "descending."

Accuracy

There are a number of ways to measure the accuracy of a model that generates probability judgments for multiple hypotheses; here we concentrate on four that we found to be particularly informative. The first two involve choosing which hypothesis is the most likely to be correct given a particular cue pattern; these choices are then compared to which hypothesis is the most likely based on the Bayesian calculations. The other two methods measure how close ESAM's probability judgments are to the probability judgments generated using the Bayesian calculations.

The simplest way to choose which hypothesis is correct for a certain cue pattern is to choose the hypothesis with the highest probability assigned to it. The expected accuracy of classifying a cue pattern will then equal the Bayesian probability of that hypothesis being the correct answer. Therefore, if the Bayesian probability of a certain hypothesis being correct given a certain set of cue values is 0.60, and ESAM assigns this hypothesis a probability of 0.65, then this method chooses this hypothesis every time and is correct on 60% of the trials involving that cue pattern. This is a *maximizing* method because choosing the hypothesis with the highest judged probability every time the cue pattern is encountered maximizes the expected proportion of correct choices.

The second method of choosing a hypothesis is *probability matching*. With this method, one uses the judged probabilities of each hypothesis as the probability of choosing that hypothesis (Luce, 1959). Therefore, if ESAM estimates the probability of hypotheses 1, 2, and 3 to be 0.65, 0.25, and 0.10 respectively, then the three hypotheses are chosen with those probabilities. If the Bayesian probabilities of the hypotheses are 0.60, 0.35, and 0.05, then the proportion of times that ESAM chooses the correct hypothesis will equal the probability of choosing each hypothesis multiplied by the probability of

the chosen hypothesis being correct, summed over all of the hypotheses. Therefore, in this example ESAM would choose the correct hypothesis with probability $(0.65 \times 0.60) + (0.25 \times 0.35) + (0.10 \times 0.05) = 0.4825$. As explained previously, if β is greater than zero then ESAM produces probability judgments that sum to more than one. In these situations, we assume that the probability of choosing each hypothesis is the probability assigned to that hypothesis divided by the sum of all of the probabilities given to that cue pattern (which normalizes the predicted probabilities to sum to one).

For both choice methods, the mean accuracy is computed by taking the expected accuracy on trials involving each cue pattern and weighting each by the proportion of times that the cue pattern would be encountered in the judgment environment (i.e., as if the cue patterns were sampled in a representative manner). The difference between these two methods is that the accuracy of the maximizing strategy depends only on the rank order of the probability judgments, whereas the accuracy of the probability matching strategy depends on the actual value of the probability judgments. Therefore, out of the two strategies the probability matching strategy is arguably more informative in the sense that it is more sensitive to the judgments produced by the model.

To identify a benchmark of ideal accuracy, the accuracy of combining all of the information in a Bayesian manner must be assessed. With the maximizing strategy, the Bayesian would always choose the hypothesis with the highest probability and would be correct with a probability equal to the Bayesian probability of that hypothesis (0.60 in our example given here). With the probability matching strategy, each category would be chosen with a probability equal to the Bayesian probability and would have that same probability of being correct (0.485 in our example). Another informative benchmark is chance-level accuracy, which can be found by computing the expected accuracy of choosing randomly among the competing hypotheses. This yields a mean accuracy of $1/N_H$ (0.333 in our example).

As can be seen in Table 12.3, these benchmarks for the maximum accuracy (Bayesian responding) and chance-level accuracy (random responding) are different in each environment, making comparisons of choice accuracy across environments difficult unless we control for this factor. We do this by calculating, for each environment, where ESAM's performance falls along the range between the chance-level and maximum choice accuracy. This is done by subtracting the chance-level accuracy from ESAM's accuracy, and dividing this by the difference between the maximum and chance-level accuracies in that environment. This yields an accuracy measure ranging between 0 and 1 that can be compared across different environments. To illustrate, in the example already discussed the Bayesian probability matching strategy resulted in an accuracy of 0.485, ESAM achieved a probability matching accuracy of 0.4825, and giving a random response would yield an

TABLE 12.3. *Accuracy Achievable by a Bayesian Responder and a Random Responder*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Bayesian Accuracy		Random Responder	
			Maximize	Probability Match	Accuracy	RMSE
Flat	3	3	0.765	0.660	0.333	0.306
		6	0.858	0.807	0.333	0.338
		12	0.956	0.937	0.333	0.373
	6	6	0.606	0.455	0.167	0.207
		12	0.748	0.656	0.167	0.239
J-shaped	3	6	0.790	0.694	0.333	0.309
		12	0.800	0.703	0.333	0.310
	6	12	0.643	0.503	0.167	0.213
Means						
Flat			0.854	0.800		
J-shaped			0.744	0.633		
	3		0.804	0.720	0.333	0.327
	6		0.666	0.538	0.167	0.220

accuracy of 0.333 on average. Therefore, the rescaled probability matching choice accuracy of ESAM would be $(0.4825 - 0.333)/(0.485 - 0.333) = 0.984$. A random responder would achieve an accuracy of 0 by this measure, whereas a Bayesian responder would achieve an accuracy of 1.

The other two methods of computing the accuracy of the probability judgments measure how close the probability judgments generated by ESAM are to the Bayesian set of probability judgments. The first of these takes all of the probability judgments and computes a correlation between the probability judgments generated by ESAM and the Bayesian set of probabilities. The second method takes the square root of the mean squared difference between each Bayesian and ESAM-generated probability, also known as a root-mean-squared error (RMSE). The RMSE is the more sensitive measure of these two, but only when ESAM's judgments for each cue pattern sum to one (i.e., when there is no subadditivity). When ESAM's judgments for each cue pattern sum to more than one then the RMSE increases accordingly. In this case, the correlation may be a more informative measure as it is insensitive to scale differences.

The bottom of Table 12.3 shows the accuracy achievable by a random and Bayesian responder for each type of environment. To compare across the two levels of one variable (e.g., flat versus J-shaped cue diagnosticity environments), the environments included in calculating a mean for one level of the variable must be identical to those included in calculating the mean for the other level of that variable. Therefore, since the J-shaped cue diagnosticity environments all have more cues than hypotheses, only the flat cue diagnosticity environments in which there are more cues than hypotheses are used to calculate the corresponding mean. This allows meaningful comparisons to be made across the variable of interest. The same logic is used in calculating the means reported in all of the tables and figures in this chapter.

ESAM's Parameters

The purpose of the simulations reported in this chapter is to assess how ESAM's performance varies based on the number of cues that are sampled in computing the probability judgments. To do this, we vary δ (the absent versus present cue weighting parameter) and β (the enhanced residual discounting parameter). All of the hypotheses have equal base rates in the environments investigated here so α (the base-rate information versus diagnostic information weighting parameter) drops out of the model.

The final parameter in ESAM is γ. Varying this affects the extremity of the probability judgments, and we have previously interpreted the value of this parameter as reflecting the amount of confidence in the judgments given (Koehler et al., 2003). This parameter has no effect on the number of cues sampled to make each judgment, and so we kept it constant across

the different simulations. Since there is no obvious default value for this parameter, we determined the value to use for all of the simulations by finding the value that allowed ESAM to fit the Bayesian probabilities as closely as possible when there was no residual discounting and absent and present cues were weighted equally ($\beta = 0$, $\delta = 0.5$).

Because the diagnostic value of each cue is summed to obtain the total diagnostic value of the cue pattern for each hypothesis (see Eq. 12.7), the absolute size of the total diagnostic value of the cue pattern is affected by the number of cues in the environment; specifically, $d_{\mathbf{C}}(H)$ varies from 0 to N_C. As a result, the value of γ that allows ESAM to best fit the Bayesian values in each environment depends on the number of cues in the environment. This complication can be avoided by changing the support calculation slightly, by exponentiating the total diagnostic value of the cue pattern $d_{\mathbf{C}}(H)$ by the number of cues in the environment N_C:

$$d_{\mathbf{C}}(H) = \left[(1-\delta) \sum_{i}^{\text{present cues}} d_1(C_i, H) + \delta \sum_{i}^{\text{absent cues}} d_0(C_i, H) \right]^{N_C}$$
$$\text{for cues } C_i,\ i = 1, \ldots, N_C, \text{ in } \mathbf{C}. \quad (12.12)$$

Using this approach, the value of γ that allows ESAM to most closely fit the Bayesian values is very close to 1.35 in each of the environments used in our simulations. We therefore used this modified form of the total cue diagnosticity equation and a value of $\gamma = 1.35$ in all of the simulations reported here.

ESAM's Performance

We begin by assessing ESAM's performance when using all of the cues, and then examine how ESAM's simplifying assumptions (which cause certain cues to be removed from the calculations) impact the model's accuracy. We first examine the impact of ESAM's simplifying assumptions by only sampling present cues, and then proceed to sample only certain subsets of the cues when evaluating support for the alternative hypotheses. Initially, we choose which cues to sample at random as in the implementation of ESAM described previously, and then adopt a more systematic method of sampling only the most diagnostic cues. We then combine these two simplifying assumptions to assess ESAM's performance when only cues of one status are sampled, and only a subset of cues are sampled when assessing support for hypotheses in the residual. It is shown that even when a large proportion of the cues are ignored, the accuracy of the choices and the probability judgments produced by ESAM remain relatively high in most of the environments examined.

We investigated the validity of ESAM's third main simplifying assumption, that all cues are assumed to be conditionally independent, in

TABLE 12.4. *Accuracy of ESAM when All Cues Are Sampled*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Probability Judgments		Choice Probability Matching (0–1 Scale)
			Correlation	RMSE	
Flat	3	3	0.999	0.016	0.979
		6	1.000	0.005	0.996
		12	1.000	0.006	1.000
	6	6	0.997	0.015	0.972
		12	0.998	0.015	0.990
J-shaped	3	6	0.996	0.027	1.008
		12	0.996	0.031	1.024
	6	12	0.996	0.020	1.018
Means					
Flat			0.999	0.009	0.995
J-shaped			0.996	0.026	1.017
	3		0.999	0.013	0.994
	6		0.997	0.017	0.993

other work (White, 2002; White & Koehler, 2003). Participants' judgments showed minor sensitivity to violations of conditional independence in environments similar to those used here, and so we extended ESAM to allow for this observed minor sensitivity. However, we leave the investigation of the impact of ignoring conditional dependencies between cues on the model's performance to future work. In the current environments, all cues were conditionally independent.

Complete Cue Sampling. ESAM's performance when all of the available information is used is remarkably good considering that the information is combined additively rather than in the mathematically correct multiplicative manner. In Table 12.4, we see that all probability matching choice accuracies exceed 0.97 and all RMSEs are 0.031 or less (using parameter values of $\beta = 0$, $\gamma = 1.35$, and $\delta = 0.5$). This performance is again consistent with previous research showing that as long as the appropriate information is used then the exact method by which it is combined is relatively unimportant (e.g., Anderson, 1981; Dawes, 1979).

We report only the probability matching choice accuracy in this and all of the subsequent analyses because the maximizing technique almost always yields performance identical to that of a Bayesian responder. This observation is interesting in itself as it shows that a vast amount of the information can be ignored and yet a good decision regarding which of the hypotheses

TABLE 12.5. *Performance of ESAM when Only Cues of One Status Are Sampled*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Probability Judgments		Choice
			Correlation	RMSE	Probability Matching (0–1 Scale)
Flat	3	3	0.987	0.050	0.985
		6	0.984	0.061	0.996
		12	0.995	0.037	0.997
	6	6	0.968	0.054	0.993
		12	0.984	0.044	0.990
J-shaped	3	6	0.972	0.076	1.011
		12	0.984	0.059	1.024
	6	12	0.977	0.049	1.024
Means					
Flat			0.988	0.047	0.994
J-shaped			0.978	0.061	1.020
	3		0.981	0.062	0.997
	6		0.976	0.049	1.002

is the *most* likely can still be made. However, as the amount of information sampled decreases, accuracy deteriorates much more quickly when evaluating the *relative* likelihood of the hypotheses compared to when only identifying which hypothesis is the most likely.

Sampling Cues of Only One Status. As already discussed, people's judgments tend to focus on the cues that are present rather than on those that are absent. Therefore, we investigated how performance decreased when only cues that are present are sampled from the available information (in the environments studied here, this is logically equivalent to only sampling cues that are absent). ESAM's performance when only cues that are present are sampled is surprisingly close to that of a Bayesian responder. See Table 12.5; here parameter values of $\beta = 0$, $\gamma = 1.35$, and $\delta = 0$ or 1 used but identical results are obtained when either extreme value of δ is used.[1] Although the probability judgments are markedly less accurate (RMSEs increased from approximately 0.01–0.02 when all cues were sampled to approximately 0.05 when only present cues were sampled), the accuracy of the choices based on those judgments remained almost unchanged.

[1] The boundary values of δ (0.0 and 1.0) cannot actually be used owing to the support for all hypotheses equaling zero when all cues have the status that is being ignored, causing the probability of each hypothesis to be undefined. Therefore, we used values of $\delta = 0.0001$ and $\delta = 0.9999$ in the simulations to avoid this.

This pattern of results is due to some of the probabilities generated being more extreme than the Bayesian values, and thereby yielding a higher accuracy measure than the Bayesian probability matching strategy, and the other probabilities being less extreme than the Bayesian values, and thereby yielding a lower accuracy measure. Specifically, when only present cues are used to make the judgments then, when less than half of the cues are present, the judgments produced are overly extreme; when more than half of the cues are present, by contrast, the judgments are less extreme than the corresponding Bayesian values. This pattern of results can best be explained with an example: If, for a given cue pattern, the Bayesian probabilities are 0.60, 0.30, and 0.10 for hypotheses #1, #2, and #3, respectively, then the Bayesian probability matching choice accuracy would be 0.46. If more extreme values are produced by ESAM of 0.75, 0.20, and 0.05, then the probability matching choice accuracy would be .515, but if less extreme values are produced of 0.45, 0.35, and 0.20 for another cue pattern that has the same Bayesian probabilities, then the accuracy would be 0.395. Therefore, on average the choice accuracy of ESAM is about the same as the Bayesian responder (means of 0.455 and 0.46, respectively, in our example), but the RMSE of ESAM's probability judgments is quite high (0.11 in our example).

This explains why the mean accuracy of ESAM when only sampling cues of one status is close to that obtained when present and absent cues receive equal weighting whereas the RMSE is greater. Therefore, we can conclude that although attending to cues of only one status may compromise the accuracy of the probability judgments generated, it may (on average) not compromise the accuracy of choices made on the basis of those judgments (i.e., identification of the most likely hypothesis).

The median best-fitting parameter value of δ for the participants making probability judgments in Experiment 3 of Koehler et al. (2003) was 0.25. When δ is set to 0.25 in ESAM in our simulations using the almost equivalent environment, the results ($\delta = 0.25$, RMSE $= 0.044$, accuracy $= 1.01$) are understandably halfway between those achieved when absent cues are ignored ($\delta = 0.0$, RMSE $= 0.076$, accuracy $= 1.01$) and when absent and present cues are weighted equally ($\delta = 0.5$, RMSE $= 0.027$, accuracy $= 1.01$). The intermediate value of δ can be viewed as representing a trade-off between the number of cues sampled and the overall accuracy afforded by the judgment strategy.

Cue Sampling When Assessing Support for the Alternatives. Descriptively, the total of the judged probabilities assigned to a set of competing hypotheses typically exceeds 1. Support theory accounts for this by positing that the support for the alternative hypotheses that are included in the residual is discounted (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997). This principle is implemented in ESAM by way of fewer cues being sampled

when computing the support for the hypotheses in the residual. We now investigate the effect that this assumption has on ESAM's performance.

The absolute support values for the hypotheses in each of the environments differ, and so the results obtained by using a fixed value of β across all of the environments are uninterpretable. Since a different proportion of cues is sampled for each judgment when assessing the support for the alternative hypotheses, the value of β was set so that the *mean* proportion of cues sampled when assessing the support for the alternative hypotheses was equal in each environment. Once Eq. 12.9 has been used to determine the proportion of cues to sample when assessing support for the alternative hypotheses, the actual cues to sample can be selected either at random or systematically. Koehler et al. (2003) assumed that the cues chosen to sample were selected randomly. Table 12.6 shows the results using this method.

There are a few interesting aspects of the results shown in Table 12.6. First, the mean total probability judgments are greater when there are six hypotheses than when there are three hypotheses. This effect has also been documented with human participants (see Tversky & Koehler, 1994, for a review), corroborating our account of the observed subadditivity. In addition, the mean total probability judgments increase as the proportion of cues sampled decreases, which is to be expected. Finally, when each judgment is based on less information (i.e., when the mean number of cues sampled to calculate the support for the alternative hypotheses decreases), the accuracy of those judgments decreases accordingly. This decrease in accuracy is seen in the correlation between ESAM's and the Bayesian probability judgments and also in the accuracy of the probability matching choices.

Reducing the number of cues sampled when assessing the support for a hypothesis will necessarily cause the judgments produced to be less accurate. However, this reduction in accuracy can be partially offset if, instead of randomly selecting which cues to sample, one chooses to sample the most diagnostic cues. To do this, an extra piece of computation is required to order the cues according to their diagnostic value. Specifically, we first order the sets of cues according to their overall diagnosticity and then within each set we order the individual cues according to how diagnostic each is for the hypothesis whose support is currently being calculated. In this way, a hierarchy of cues can be created. The appropriate proportion of cues to sample is then determined by starting with the cue highest on this ordering and including all cues down to a level that achieves the desired sampling proportion. This general process is similar to that used by the "Take the Best" algorithm (Gigerenzer & Goldstein, 1999), which orders the cues in terms of diagnosticity and then makes a decision based exclusively on the most diagnostic cue for which the two objects currently being assessed have different values.

TABLE 12.6. *Performance of ESAM when Only a Certain Proportion (a Random Subset) of the Cues Is Sampled to Compute the Support for the Alternative Hypotheses*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Proportion Sampled	Probability Judgments						Choice		
				Correlation			Total Probability Judgment			Probability Matching (0–1 scale)		
				0.9	0.8	0.6	0.9	0.8	0.6	0.9	0.8	0.6
Flat	3	3		1.00	0.99	0.83	1.18	1.35	1.82	0.96	0.88	0.61
		6		0.98	0.95	0.72	1.23	1.41	1.96	0.98	0.92	0.66
		12		0.97	0.91	0.67	1.22	1.41	1.91	0.98	0.94	0.75
	6	6		0.96	0.85	0.65	1.76	2.40	3.56	0.87	0.64	0.36
		12		0.89	0.76	0.56	1.92	2.60	3.71	0.75	0.52	0.28
J-shaped	3	6		0.97	0.92	0.75	1.29	1.53	1.98	0.97	0.88	0.60
		12		0.95	0.85	0.64	1.42	1.74	2.23	0.95	0.78	0.45
	6	12		0.89	0.76	0.55	2.08	2.85	4.04	0.76	0.52	0.27
Means												
Flat				0.95	0.87	0.65				0.90	0.79	0.56
J-shaped				0.94	0.84	0.65				0.89	0.73	0.47
	3			0.98	0.95	0.77	1.23	1.43	1.92	0.97	0.89	0.62
	6			0.92	0.79	0.59	1.92	2.62	3.77	0.79	0.56	0.30
Mean				0.95	0.87	0.57				0.90	0.76	0.42

P1: IYP
0521831598c12.xml CB873/Fiedler 0 521 83159 8 August 22, 2005 18:37

The measure of cue diagnosticity used to establish this ordering relies on computations already carried out by ESAM. Specifically, given a particular hypothesis the diagnosticity of a cue being present and being absent is calculated using the appropriate version of Eq. 12.5. The cue's diagnosticity for that particular hypothesis is defined as the absolute difference between these two values. The cue's overall diagnosticity for the full set of hypotheses is then calculated as the mean of these absolute differences across the set of hypotheses.

Table 12.7 shows ESAM's performance using this systematic cue sampling method. Qualitatively, the results are similar to those produced when randomly selecting which cues to sample. The main difference is that the decrease in accuracy resulting from a reduction in the number of cues sampled is markedly less dramatic using the systematic cue sampling method than it is using the random cue sampling method. This difference is shown in Figure 12.2, in which the mean correlation and choice accuracy across all of the environments is plotted for each method of cue selection. This finding is to be expected since the most diagnostic information is sure to be retained when the cues are selected based on their diagnosticity whereas the most diagnostic information will often be discarded when the cues are selected randomly.[2]

As long as at least 90% of the cues (approximately) are sampled on average when assessing the support for the alternative hypotheses (selecting the subset of cues using the systematic method), the probability matching choice accuracy is maintained at around the maximum level of 1; see Figure 12.3. Below 90%, the accuracy starts to fall off reasonably quickly in most of the environments. The main reason for this decrease is that if the average proportion of cues sampled falls to lower levels, then if the focal hypothesis has at least a moderate amount of support then very few cues are sampled when assessing the support for the alternatives, thereby causing the focal hypothesis to receive a high probability judgment. This reduces accuracy because if two or more hypotheses have a moderate amount of support they will all be given a very high probability judgment, and so the model loses the power to differentiate between these hypotheses. Therefore, ESAM

[2] Comparing Tables 12.6 and 12.7 shows that selecting the cues to sample systematically yields lower mean total probability judgments relative to when the selection is done randomly. This is due to the alternative hypotheses not losing as much support when the sampling is done systematically because the cues that may contribute most to the support of the alternatives are certain to be included in the calculation of the support for those hypotheses. This contrasts with the random selection method in which the same cues are no more likely to be selected than are cues that contribute less to the support. Therefore, on average the support for the alternative hypotheses is greater when the cues are selected systematically than when selected randomly, despite the fact that the same proportion of cues is sampled with both methods.

TABLE 12.7. *Performance of ESAM when Only a Certain Proportion (a Systematic Subset) of the Cues Is Sampled to Compute the Support for the Alternative Hypotheses*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Proportion Sampled	Probability Judgments						Choice		
				Correlation			Total Probability Judgment			Probability Matching (0–1 scale)		
				0.9	**0.8**	**0.6**	**0.9**	**0.8**	**0.6**	**0.9**	**0.8**	**0.6**
Flat	3	3		1.00	0.92	0.93	1.11	1.55	1.58	1.02	0.79	0.81
		6		0.97	0.97	0.80	1.18	1.18	1.80	1.01	1.01	0.79
		12		1.00	0.93	0.69	1.02	1.24	1.87	1.00	0.97	0.80
	6	6		0.94	0.84	0.73	1.56	2.30	3.07	1.02	0.70	0.48
		12		0.87	0.77	0.58	1.64	2.16	3.47	0.82	0.63	0.34
J-shaped	3	6		0.97	0.94	0.73	1.22	1.44	1.97	1.02	0.92	0.62
		12		0.94	0.88	0.64	1.42	1.65	2.21	0.95	0.84	0.48
	6	12		0.87	0.75	0.57	1.92	2.75	3.85	0.84	0.55	0.30
Means												
Flat				0.95	0.89	0.69				0.94	0.89	0.64
J-shaped				0.93	0.86	0.65				0.94	0.77	0.47
	3			0.98	0.94	0.82	1.17	1.39	1.78	1.02	0.90	0.74
	6			0.89	0.79	0.63	1.71	2.40	3.46	0.89	0.63	0.37
Mean				0.94	0.88	0.71				0.96	0.80	0.58

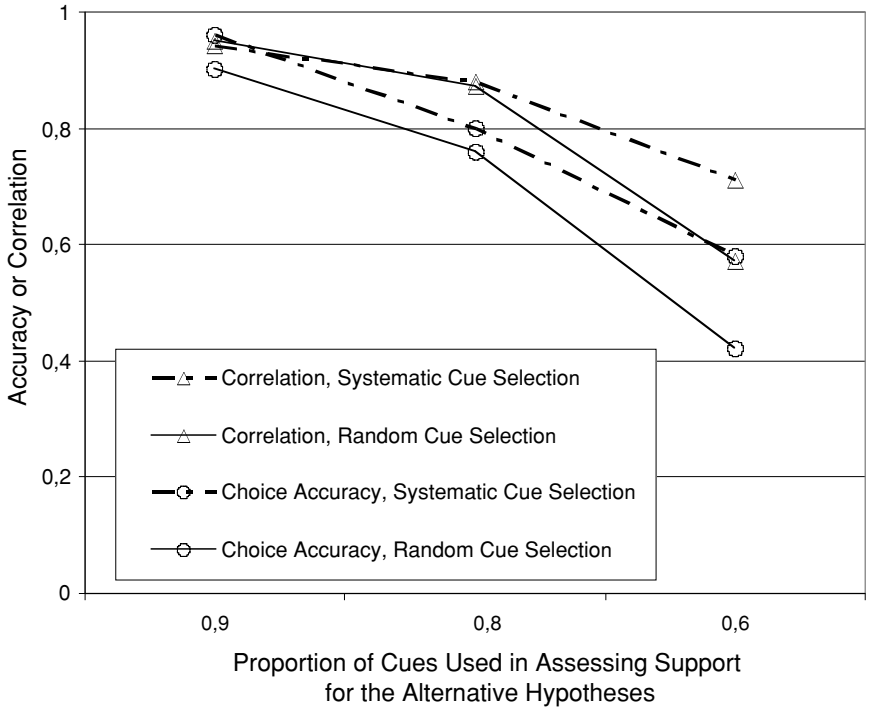

FIGURE 12.2. Random versus systematic selection of cues used in assessing support of the alternative hypotheses.

chooses each hypothesis that has a reasonable amount of support with almost equal frequencies.[3]

The median value of the observed mean total probability judgment in Experiment 3 of Koehler et al. (2003) was 1.31. When the value of β is set such that ESAM produces a mean total probability judgment of this size (given a similar environment) then the cue sampling proportion is 0.87, and the probability matching choice accuracy is 0.96. This implies that the number of cues sampled to compute the support for the alternative hypotheses, at least in this study, is set about as low as possible without sustaining any significant decrement in accuracy.

Cue Sampling When Assessing Support for the Alternatives and Sampling Cues of Only One Status. Because of its greater accuracy, we use the systematic method of cue sampling to assess ESAM's performance when both simplifying assumptions are implemented concurrently. Surprisingly,

[3] This is basically a problem in the extremity of the support values, and so may be affected by manipulating the γ parameter.

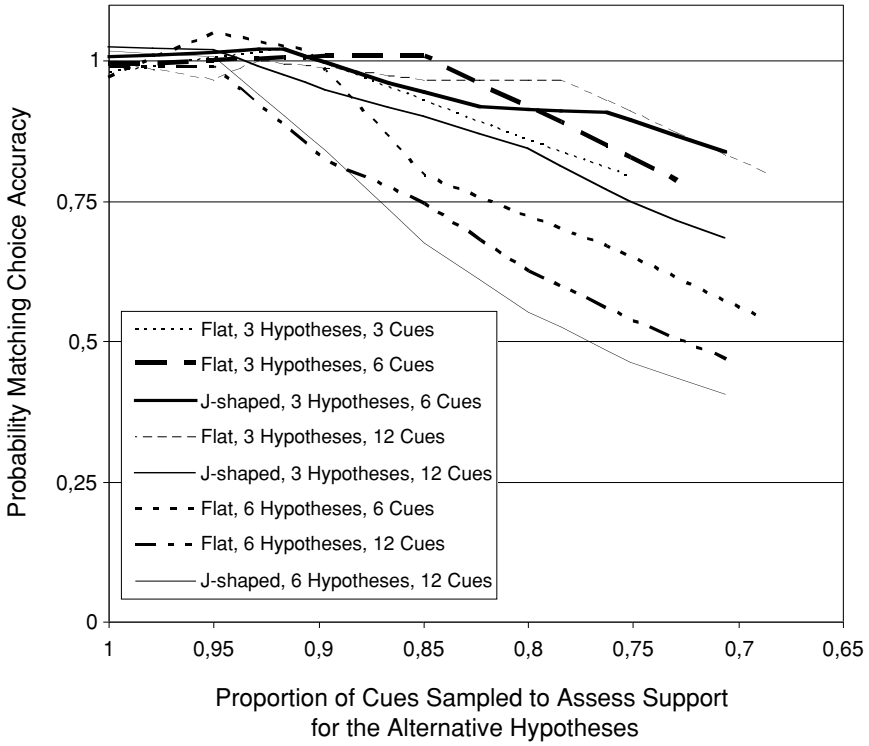

FIGURE 12.3. Probability matching choice accuracy versus proportion of cues sampled when assessing support for the alternative hypotheses.

although the correlation between ESAM's probability judgments and the Bayesian values decreases in comparison to the results of the simulations reported in the foregoing, the choice accuracy increases (see Table 12.8; the random method of cue selection yields a qualitatively similar pattern of results). The reason is again that by only sampling cues of a certain status the judgments produced become more extreme, causing the probability matching choice strategy to more closely resemble the maximizing choice strategy whereas the judgments themselves diverge from the Bayesian values more.

We can again investigate how the model performs when parameter values are used that allow the model to mimic the human data from Experiment 3 of Koehler et al. (2003). If values of $\delta = 0.25$ and a systematic cue sampling proportion of 0.88 are used (mean total probability judgment = 1.31), then the probability judgments generated correlate quite highly with the Bayesian probabilities ($r = 0.882$) and choice accuracy is reasonably high (0.935). If we assume that a value of $\delta = 0.25$ means that all of the present cues are sampled, and one-third of the absent cues are sampled

TABLE 12.8. *Performance of ESAM when Only a Certain Proportion (a Systematic Subset) of the Cues Is Sampled to Compute the Support for the Alternative Hypotheses, and Only Cues of a Certain Status within Those Sampled Are Used*

Cue Diagnosticity	Number of Hypotheses	Number of Cues	Proportion Sampled	Probability Judgments						Choice		
				Correlation			Total Probability Judgment			Probability Matching (0–1 scale)		
				0.9	0.8	0.6	0.9	0.8	0.6	0.9	0.8	0.6
Flat	3	3		0.92	0.75	0.81	1.16	1.46	1.57	0.99	0.80	0.83
		6		0.89	0.81	0.66	1.21	1.39	1.78	0.98	0.85	0.80
		12		0.85	0.79	0.63	1.23	1.35	1.79	0.94	0.90	0.77
	6	6		0.75	0.64	0.50	1.59	2.21	3.05	0.99	0.90	0.62
		12		0.72	0.60	0.47	1.55	2.06	3.05	0.92	0.79	0.60
J-shaped	3	6		0.83	0.74	0.60	1.29	1.49	1.83	0.94	0.86	0.74
		12		0.80	0.70	0.55	1.32	1.54	1.93	0.93	0.84	0.62
	6	12		0.66	0.55	0.44	1.74	2.35	3.23	0.94	0.80	0.59
Means												
Flat				0.82	0.73	0.59				0.95	0.85	0.72
J-shaped				0.76	0.66	0.53				0.94	0.83	0.65
	3			0.88	0.77	0.69	1.22	1.45	1.73	0.97	0.84	0.79
	6			0.71	0.60	0.47	1.63	2.21	3.11	0.95	0.83	0.60
Mean				0.80	0.70	0.58				0.95	0.84	0.70

($\delta/[1-\delta] = 0.25/0.75 = 1/3$) then the model uses only 62% of the cue frequency counts used by the full Bayesian method, and yet it loses very little in terms of accuracy.

Conclusions

We have shown that when all of the available information characterizing previous observations in an uncertain environment is used, ESAM can produce choices and probability judgments that are remarkably close to those obtained by using the normatively correct Bayesian calculations. This is despite ESAM combining the information from the individual cues for each hypothesis additively rather than multiplicatively.

Interestingly, the accuracy of the choices made does not decrease when only cues of a certain status are sampled. At the same time, the probability judgments themselves are slightly less accurate, but this can be explained by the probabilities for certain cue patterns becoming more extreme whereas the probabilities for others become more conservative when this method of sampling is used. We also investigated the accuracy of the responses generated by ESAM when only some of the cues are sampled to assess the support for the alternative hypotheses. The decreases in accuracy were less pronounced when the cues not sampled were the least diagnostic cues rather than a random subset of the cues. It was shown that a small proportion of the cues can be ignored when assessing the support for the alternative hypotheses without any decrease in choice accuracy, whereas the total probability judgments for each cue pattern are significantly greater than one. When ESAM generates probability judgments that show biases similar to those exhibited by the participants in Experiment 3 of Koehler et al. (2003), only 62% of the cues are sampled to generate probability judgments that are highly accurate.

This research demonstrates how, when only a subset of the information available is sampled, reasonably accurate responses can still be generated. Furthermore, we have shown that when ESAM displays biases of a similar magnitude to those exhibited by humans (with the consequence that far less information is sampled than is available), the accuracy of the probability judgments and choices generated by ESAM is still close to that obtainable when all of the information is sampled.

OTHER MODELS

Other research investigating sampling subsets of cues includes Gigerenzer and Goldstein's (1999) "Take the Best" heuristic. They found that accurate choices could be made when only one cue or a few of the cues were considered when making each choice. Here, we have extended this approach of selectively sampling certain cues but used a quite different task

with very different methods of choosing which cues to sample. However, we reached the same conclusion: Not all of the information needs to be sampled to make a reasonably accurate choice or judgment.

In evaluating the diagnostic value of a cue pattern, ESAM integrates separate assessments of the implications of each individual cue value constituting the cue pattern, and hence it implicitly assumes conditional independence of cues. Effectively, then, ESAM can be viewed as employing a prototype representation of cue information, in which the interpretation of each cue value is uninfluenced by other cue values in the cue pattern. Exemplar-based models, by contrast, have the potential advantage of detecting and exploiting configural information in cases where conditional dependencies exist among cue values, but at the cost of substantially higher memory storage requirements. For example, in the case of six binary cues (i.e., symptoms) and three competing hypotheses as investigated in many of our experiments, counts must be maintained of the frequency with which each of the sixty-four possible cue patterns co-occur with each hypothesis in an exemplar representation. The prototype representation requires only twelve counts (i.e., an absence and a presence count for each cue) per hypothesis, rather than sixty-four, because it maintains counts for each separate cue value rather than for entire cue patterns. In essence, the prototype representation as employed by ESAM discards information that the exemplar representation retains.

Other researchers have studied ways in which entire exemplars could be selectively sampled. One instance of this is Dougherty et al.'s (1999) extension of Hintzman's (1988) MINERVA-2 memory model to decision-making tasks (MINERVA-DM), and in particular to judgments of conditional likelihood. The output of MINERVA-DM is based on imperfect exemplar representations (or memory traces). In one simulation, they provided three sets of simulated participants different amounts of experience with a domain, by storing 80, 200, or 600 memory traces, each representing a patient displaying ten present/absent symptoms that probabilistically predicted each of two diseases. This manipulation is functionally equivalent to storing all 600 traces and subsequently sampling 13.3%, 33.3%, or 100% of them. The resulting judgments exhibited less overconfidence when they were based on a larger amount of experience. The same pattern of results has been reported with human participants in multiple-cue judgment tasks (see Dougherty et al., 1999, for a discussion of the human data). However, again in line with the results presented in this chapter, the large differences in the amount of information sampled resulted in relatively small differences in the measure they used (overconfidence). However, Dougherty et al. (1999) did not report any direct measures of the accuracy of the judgments produced by their model.

Juslin and Persson (2002) suggested a modification of Medin and Schaffer's (1978) context model in which all exemplars are activated in parallel

but are retrieved, or sampled, serially (PROBEX – PROBabilities from EXemplars). The order of retrieval of the exemplars is determined by the most similar exemplar to the current probe being the most likely to be sampled first. Once "a clear-enough conception of the estimated quantity has been attained" (p. 569) retrieval terminates. In addition, although the status of all the cues is known when each exemplar is encountered, the status of each cue in each exemplar has a certain probability of not being available upon retrieval due to an encoding failure or forgetting. This is functionally equivalent to sampling only a subset of the cues for each exemplar.

PROBEX was able to mimic human performance by only sampling a few exemplars before making a response. The researchers also claimed that the model was robust even when only a few exemplars are stored (Juslin, Nilsson, & Olsson, 2001; Juslin & Persson, 2002). However, they did not directly measure changes in performance over varying amounts of sampled information. Instead, it was suggested that

> an important research program [would be] 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? (p. 598, 2002)

Given the interesting findings presented here regarding the effects of limited sampling in a cue-frequency-based model, ESAM, we agree with Juslin & Persson's (2002) suggestion that similar research should be conducted with an exemplar-based model.

SUMMARY

To make a reasonably accurate prediction or diagnosis in an uncertain environment on the basis of a set of informational cues and previous experience, not all of the available information needs to be sampled nor does it need to be combined in the mathematically correct way. We assessed the accuracy of a mathematical model of human probability judgment, the Evidential Support Accumulation Model (ESAM), in a variety of environments. ESAM is a model of how the evidence from multiple cues is combined to assess the support for conflicting hypotheses, with the evidence from each cue being combined additively. Derived from the principles of support theory, the support for the focal hypothesis is assessed by sampling more cues than when assessing the support for the alternative hypotheses, thereby mimicking the subadditivity observed in human probability judgments. In addition, ESAM assumes that cues of different statuses (e.g., present versus absent) that reflect differences in psychological salience are sampled with differing probabilities. In the eight simulated environments studied, when only a subset of the available information was sampled only

a small decrement in the judgmental accuracy achieved by the model was observed.

References

Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

Brooks, L. R. (1978). Non-analytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169–211). Hillsdale, NJ: Lawrence Erlbaum.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dougherty, M. R. P., Gettys, C. E., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review, 106* (1), 108–209.

Edgell, S. E. (1978). Configural information processing in two-cue probability learning. *Organizational Behavior & Human Decision Processes, 22*, 404–416.

Edgell, S. E. (1980). Higher order configural information processing in nonmetric multiple-cue probability learning. *Organizational Behavior & Human Decision Processes, 25*, 1–14.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception & Performance, 4*, 330–334.

Gigerenzer, G., & Goldstein, D. G. (1999). Betting on one good reason: The take the best heuristic. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart: Evolution and cognition*. London: Oxford University Press.

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

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: Lawrence Erlbaum.

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26*, 563–607.

Koehler, D. J. (2000). Probability judgment in three-category classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 28–52.

Koehler, D. J., Brenner, L. A., & Tversky, A. (1997). The enhancement effect in probability judgment. *Journal of Behavioral Decision Making, 10*, 293–313.

Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology, 46*, 152–197.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart: Evolution and cognition*. London: Oxford University Press.

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

Novemsky, N., & Kronzon, S. (1999) How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. *Journal of Behavioral Decision Making, 12*, 55–69.

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*, 406–415.

Sedlmeier, P., & Betsch, T. (2002). *ETC. Frequency processing and cognition*. Oxford: Oxford University Press.

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review, 101*, 547–567.

Wasserman, E. A., Dorner, W. W., & Kao, S. F. (1990). Contributions of specific cell information to judgments of inter-event contingency. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*, 509–521.

White, C. M. (2002). *Modelling the influence of cue dependencies and missing information in multiple-cue probability learning*. Master's thesis, University of Waterloo, Waterloo, Ontario, Canada.

White, C. M., & Koehler, D. J. (2003). *Modeling the influence of cue dependencies in multiple-cue probability learning*. Unpublished manuscript, University of Waterloo, Waterloo, Ontario, Canada.