ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cognitive Psychology 46 (2003) 152–197

Cognitive Psychology

www.elsevier.com/locate/cogpsych

An evidential support accumulation model of subjective probability☆

Derek J. Koehler,* Chris M. White, and Ray Grondin

Department of Psychology, University of Waterloo, Waterloo, Ont., Canada N2L 3G1

Accepted 8 May 2002

Abstract

A model of cue-based probability judgment is developed within the framework of support theory. Cue diagnosticity is evaluated from experience as represented by error-free frequency counts. When presented with a pattern of cues, the diagnostic implications of each cue are assessed independently and then summed to arrive at an assessment of the support for a hypothesis, with greater weight placed on present than on absent cues. The model can also accommodate adjustment of support in light of the baserate or prior probability of a hypothesis. Support for alternatives packed together in a ''residual'' hypothesis is discounted; fewer cues are consulted in assessing support for alternatives as support for the focal hypothesis increases. Results of fitting this and several alternative models to data from four new multiple-cue probability learning experiments are reported.

Keywords: Evidence evaluation; Subjective probability; Multiple-cue probability learning; Support theory; Bayesian reasoning

1. Introduction

Support theory (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) provides a general framework for development of descriptive models of subjective

☆ This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to the first author. We are grateful to Natasha Chlebowsky and Anya Rameshwar for their assistance in conducting the experiments, and to Lyle Brenner for helpful comments.

* Corresponding author.
E-mail address: dkoehler@uwaterloo.ca (D.J. Koehler).

0010-0285/03/$ - see front matter
doi:10.1016/S0010-0285(02)00515-7

probability. As elaborated below, support theory represents judged probability as the balance of evidential support for the focal hypothesis relative to that for its alternative. While support theory has been used to successfully predict a number of judgmental phenomena, particularly effects of unpacking a hypothesis into its components, it does not specify how the perceived support for a hypothesis is determined by the available evidence. The support for a hypothesis may be based on "natural assessments" (Tversky & Kahneman, 1983) of availability or representativeness, or on more effortful assessments of arguments, frequency estimates, or probabilistic cues.

In other words, the processes underlying the assessment of evidential support will typically vary across different judgment tasks. A single model is unlikely to be able to capture all of these diverse processes at a satisfactory level of detail. Support theory thus provides a general framework within which to develop models of evidence evaluation that specify how support is assessed in a particular task or domain. In the present research, we develop such a model to describe how a set of binary (present/absent) cues is evaluated in arriving at an assessment of the extent to which a cue pattern supports a designated hypothesis, where the perceived diagnostic value of the cues is based on previous experience. Under these circumstances, we assume that support is evaluated on the basis of the experienced frequencies with which each cue value has been observed to co-occur with the outcome or hypothesis of interest. As noted above, in other tasks and domains, assessments of evidential support may be driven less by previously observed frequencies and more by other considerations such as the representativeness of the outcome with respect to the evidence at hand.

Multiple-cue probability learning studies have a long tradition in the study of human judgment (see Castellan, 1977, for a review of early work), interest in which is generally attributed to the influential work of Brunswik (1956). This tradition focuses on judgments made in an uncertain environment characterized by cues that are probabilistically related to an outcome variable of interest to the judge. The standard example is that of a physician attempting to diagnose a patient's illness (the outcome variable) on the basis of a set of diagnostic but less than perfectly predictive symptoms (the cues). In multiple-cue probability learning experiments, participants learn the predictive value of the cues on the basis of "experience" in the form of exposure to a series of individual cases (i.e., training trials) for which they are provided with information regarding the outcome and the associated cue values. Our focus is on judgments of probability following such experience, in which participants rely on what they have learned to judge the probability of a target outcome given a particular pattern of cues (e.g., judging the likelihood that a patient has a particular illness on the basis of the symptoms exhibited by that patient).

There is substantial evidence that people can accurately encode and later reproduce the frequency with which they have been exposed to the occurrence of various events (e.g., Hasher & Zacks, 1984). Likewise, there is considerable evidence that people can often identify diagnostic cues that are useful for predicting an event's occurrence on the basis of such experience (e.g., Klayman, 1988; Trope & Mackie, 1987). Finally, research also indicates that people can integrate the implications of available diagnostic cues in arriving at what are often reasonably accurate assessments of the associated

outcome's probability (e.g., Estes, 1976; Peterson, Hammond, & Summers, 1965). In light of these observations, a number of models have been developed to account for how people make such judgments and to predict when their judgments will be accurate or subject to systematic bias (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Nosofsky, Kruschke, & McKinley, 1992). As elaborated below, however, the support theory framework allows development of a model that can account for certain aspects of such judgments which are not easily accommodated by previous models.

Understanding judgments based on probabilistic cues is useful because they reflect a fundamental way in which people rely on their past experience in making predictions about future outcomes. Judgments based on learning from previous experience with probabilistic cues play a critical role in many different types of decisions, including, for example, consumer purchases (Hutchinson & Alba, 1991; Meyer, 1987; Van Osselaer & Alba, 2000). Such experience may constitute an important component of expertise in some domains, such as medicine, where diagnostic reasoning is heavily influenced by previous cases observed by the physician (e.g., Brooks, Norman, & Allen, 1991). Development of descriptive models that accurately capture cue-based judgments, then, could benefit those who wish to understand or improve judgments and decisions made under conditions of uncertainty.

In this paper we develop and test one such model using the support theory framework. In Section 1, we review the key aspects of support theory relevant to the present treatment. In Section 2, we describe previous findings from multiple-cue probability learning studies by Koehler (2000) that guided development of the present model. In Section 3, we describe the new model, which we refer to as an Evidential Support Accumulation Model (or ESAM) for reasons that will become clear later. In Section 4, we provide an overview of four new cue learning experiments. In Section 5, we use these new datasets to fit ESAM and then to compare its performance on a feature-by-feature basis with alternative models as a way of providing some corroboration for each of ESAM's critical assumptions. In Section 6, we develop a variant of ESAM that generalizes the Bayesian model. We conclude by discussing related work and some issues that we suggest merit further investigation.

2. Overview of support theory

In contrast to probability theory, in which probability is assigned directly to set-theoretic events, support theory assigns probability to descriptions of events, referred to as *hypotheses*. Support theory is thus nonextensional, allowing different probability judgments to be assigned to different descriptions of the same event. This complication is necessary to accommodate the observation that people's 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:

$$P(A,B) = \frac{s(A)}{s(A) + s(B)}. \tag{1}$$

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 distinguishes two kinds of hypotheses: explicit disjunctions, which list their components, and implicit disjunctions, which do not. 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

$$s(A) \leqslant s(A_1 \vee A_2) \leqslant s(A_1) + s(A_2). \tag{2}$$

That is, the support of the implicit disjunction A is less than or equal to that of the explicit disjunction $A_1 \vee A_2$, 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

$$T = P(A,\overline{A}) + P(B,\overline{B}) + P(C,\overline{C}) \tag{3}$$

assigned to the three elementary hypotheses will generally exceed one, in violation of probability theory. The degree of subadditivity in the elementary judgments can be measured by the extent to which the total probability T assigned to them exceeds

one; the greater the value of T, the greater the degree of subadditivity. One of the goals of the present research is to develop a model that can account for variance in T, that is, in the degree of subadditivity exhibited in a set of elementary judgments.

One drawback of this measure is that T is an aggregate representation of the general degree of subadditivity across a set of elementary probability judgments. 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 $\overline{A}$:

$$s(\overline{A}) = w_{\overline{A}}[s(B) + s(C)]. \tag{4}$$

Support theory's assumption of subadditivity (2) implies $w_{\overline{A}} \leqslant 1$. 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 $\overline{A}$.

Combining Eqs. (1) and (4), we see that the elementary probability judgment with A as the focal hypothesis is given by

$$P(A, \overline{A}) = \frac{s(A)}{s(A) + w_{\overline{A}}[s(B) + s(C)]}. \tag{5}$$

This form illustrates the advantage conveyed to the focal hypothesis in the elementary judgment, in which it receives the full support accorded to it by the evidence, while the support for its alternatives is discounted as a result of their being packed in the residual.

In later work, 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:

$$w_{\overline{A}} = 1 - \beta s(A), \tag{6}$$

where $\beta \geqslant 0$ is a free parameter. This model captures the intuition that when the focal hypothesis is well supported by the available evidence, people are generally less willing 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*. The model developed in this paper incorporates the assumption of enhanced residual discounting, though it is implemented in a somewhat different manner than in previous research.

3. Support assessment in cue-based judgment: Review of previous findings

Koehler (2000) reported the results of several initial experiments investigating the evidential determinants of subadditivity in cue-based probability judgments. Participants played the role of a physician in a simulated medical diagnosis task. They were presented with "patients" each known to be suffering from one of three possible flu strains. Participants were presented with a series of patients, each characterized in

terms of discrete binary-valued symptoms (i.e., cues) known to be present or absent (e.g., *cough*, *headache*, and *sore throat*). In a training phase, participants indicated which of the three flu strains they believed the patient to be suffering from, immediately followed by outcome feedback revealing which flu strain the patient actually had. In this manner, participants eventually learned about the predictive relationship between the symptoms and flu strains. Following the training phase, participants were presented with a further series of patients (without feedback), and for each judged the elementary probability that the patient was suffering from a designated flu strain rather than from either of its two alternatives. Three elementary judgments (one for each flu strain), separated from one another by a number of intervening items, were elicited for every possible pattern of symptoms. As predicted by support theory, these judgments were systematically subadditive, as indicated by their sum $T > 1$.

This design allows investigation of how the degree of subadditivity observed in the elementary judgments (as measured by T) varies as a function of the pattern of symptoms or cues serving as the evidential basis of the judgment. Consider as an example an experiment (Koehler, 2000, Experiment 3) that involved five symptoms. Three of the symptoms were diagnostic with respect to the patient's flu strain: The presence of a given diagnostic symptom was probabilistically associated with the presence of a corresponding flu strain, with one diagnostic symptom mapped in this fashion onto each of the three possible flu strains. The remaining two symptoms were nondiagnostic, that is, they co-occurred equally often with each of the three flu strains. The nondiagnostic symptoms differed in their overall prevalence, with one being present for 75% of all patients and the other being present for only 25% of all patients.

Three characteristics of a given symptom pattern were examined as possible influences on the degree of subadditivity T for the set of elementary judgments associated with that symptom pattern. First, the value of T increased with the number of flu strains implicated by the symptom pattern, as measured by the number of diagnostic symptoms present in the pattern. Second, the value of T also increased with the number of nondiagnostic symptoms present in the symptom pattern. The overall prevalence of a symptom (i.e., how frequently it was observed among patients in the training phase), which was the third characteristic considered, had no influence on the value of T.

These results provide an initial glimpse into the process by which cue-based evidence is evaluated in assessing the support for a hypothesis. Based on these results, Koehler (2000) offered a tentative set of principles governing the support assessment process, which are briefly reviewed here as they constitute the basis for development of the current model. These principles presuppose support theory's basic assumptions and also rely on the enhanced residual discounting assumption underlying the linear-discounting model of Koehler et al. (1997), according to which greater values of T are associated with bodies of evidence that provide a generally greater level of support for each in the set of hypotheses under consideration. Adherence to the principles outlined below simplifies the computational and memory requirements of the support assessment process, but at the

price of ignoring potentially useful information and deviating from the normative requirements of the Bayesian approach.

(A) *Composite residual formation*. As implied by support theory, alternatives to the focal hypothesis of the elementary judgment are packed together and evaluated as a single entity, losing support in the process. The observation of $T > 1$ is consistent with this claim.

(B) *Evidence decomposition*. Rather than assessing the implications of a pattern of cues taken as a whole, each cue's contribution to the support for a hypothesis is assessed individually. The observation that cue patterns with identical diagnostic implications according to the Bayesian approach (such as the pattern in the experiment described above with all three diagnostic symptoms present versus the pattern with all three absent) produce systematically different values of T is consistent with this claim.

(C) *Cue presence/absence asymmetry*. Given binary cues representing the presence or absence of a feature, the support for a hypothesis appears to be determined primarily by the present cues constituting the cue pattern. The observation that T increases with the number of present cues in the cue pattern is consistent with this claim.

(D) *Noncompensatory support assignment*. The support for a hypothesis reflects only those aspects of the evidence that directly implicate that hypothesis; evidence that implicates an alternative hypothesis increases support for the alternative hypothesis but does not directly decrease support for the non-implicated hypothesis. The observation that T increases with the number of implicating cues is consistent with this claim.

(E) *Diagnosticity-based support assignment*. The support assessment process appears to be sensitive to the diagnosticity of individual cues, such that the presence of diagnostic cues increases the support for the implicated hypothesis. The observation that mere cue prevalence in the absence of any diagnostic value has no influence on T is consistent with this claim.

(F) *Support accumulation*. The assessment of evidence can be characterized as one in which positively valued support is accumulated over the individual pieces of evidence (i.e., cues) as the body of evidence is evaluated. The observation that even the presence of nondiagnostic cues in the cue pattern produces greater values of T is consistent with this claim.

Although postulation of these principles goes well beyond the empirical results of the Koehler (2000) study, they provide a characterization of the support assessment process that can serve as the basis for development of a mathematical model. According to this characterization, when assessing the support for a hypothesis conveyed by evidence in the form of a cue pattern, the evidence is assessed one cue at a time with a focus primarily on present rather than absent cue values. Each present cue adds to the support of a hypothesis (to a greater extent than do absent cues), with greater support being added the more diagnostic the cue is with respect to the hypothesis in question. In the next section, we develop a model with these characteristics that predicts the support for a hypothesis on the basis of previous observations of the frequency with which the relevant cues co-occur with the hypotheses (i.e., outcomes) of interest.

4. Evidential support accumulation model

Support theory describes the translation of support into probability but, as discussed in the introduction, does not specify how support is assessed in the evaluation of the available evidence. Here we develop a model of the support assessment process underlying judgments of probability based on patterns of binary (present/absent) cues that follows the principles of support assessment offered in the previous section. This model, called ESAM (for Evidential Support Accumulation Model), is a work in progress that may be subject to revision in light of further testing. Furthermore, we suspect that other models that behave in accord with the principles outlined above are likely to perform comparably in terms of fit to people's judgments. ESAM is intended primarily as a demonstration of the usefulness of the support theory framework in the development of models of evidential support assessment tailored to specific judgment tasks.

4.1. Stored frequency counts

ESAM specifies how a set of observations represented in the form of frequency counts regarding cue and hypothesis co-occurrence is used to assess the support for a hypothesis provided by a particular cue pattern. For simplicity, we assume that these frequency counts, obtained from experience in the probabilistic cue-based environment, are encoded and later retrieved without error. A more sophisticated version of the model might relax this assumption by incorporating established principles of memory that provide a more accurate representation of what people are actually likely to recall from their previous experience. Estes (1986), for example, explores a family of such array-based models and considers the implications of information loss for subsequent judgments based on the imperfect memory array. Later in this paper we do consider the consequences of imperfect memory for previous observations, and find that the assumption of error-free frequency counts does not appear to detract substantially from the model's performance in our experiments. Nonetheless, it should be said that our model is not directly concerned with the process by which cue frequencies are learned, but rather with the process by which learned frequencies are used in assessing evidential support in inferential judgments. (For models of learning in multiple-cue environments, see Gluck & Bower, 1988; Kruschke & Johansen, 1999; Nosofsky et al., 1992.)

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 illustrate the model with the case of three possible hypotheses and six binary cues. This case corresponds to new experiments, reported below, 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 given 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).

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 $f_0(C,H)$. For example, if the judge had observed 12 patients with a particular flu strain for whom a specific symptom was present and six more patients with the same flu strain for whom the symptom was absent, then $f_1 = 12$ and $f_0 = 6$. To predict every possible elementary judgment (i.e., the judged probability of each hypothesis given every possible cue pattern) in the general case of N_H hypotheses and N_C binary cues, the model requires $2N_HN_C$ such frequency counts as represented in a frequency table with $N_H(N_C + 1)$ degrees of freedom. Later we also consider a simpler version of the model that relies only on counts of cue presence (i.e., one that entirely ignores cue absence) and consequently requires only half this number of values as input.

For purposes of exposition, it is useful to define the marginal frequencies with which each hypothesis holds and each cue is either present or absent, as follows. Let

$$f_1(C) = \sum_j f_1(C,H_j) \text{ over hypotheses } H_j, \quad j = 1,\ldots,N_H \tag{7a}$$

represent the overall frequency with which cue C is observed as being present in the set of stored observations, and likewise let

$$f_0(C) = \sum_j f_0(C,H_j) \tag{7b}$$

represent the overall frequency with which cue C is observed as being absent. Finally, let $f(H)$ represent the overall frequency with which hypothesis H holds in the set of stored observations. Note that in the case of binary cues,

$$f(H) = f_0(C_i,H) + f_1(C_i,H) \text{ for any cue } C_i, \quad i = 1,\ldots,N_C. \tag{8}$$

This value gives the "baserate" frequency with which hypothesis H holds in the set of stored observations.

4.2. Diagnostic implication of a cue value for a hypothesis

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 baserate 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:

$$d_1(C,H) = \frac{f_1(C,H)/f(H)}{\sum_j [f_1(C,H_j)/f(H_j)]}. \tag{9a}$$

The value of d varies between 0 and 1. If the presence of cue C is nondiagnostic with respect to the hypothesis H, then $d_1(C,H) = 1/N_H$. If the co-occurrence of cue C's presence with hypothesis H is more frequent in the set of observations than would be expected on the basis of H's baserate frequency $f(H)$ alone, then $d_1(C,H) > 1/N_H$. If the co-occurrence of cue C's presence with hypothesis H is less frequent than would be expected on the basis of H's baserate frequency, then $d_1(C,H) < 1/N_H$. This calculation can be thought of as involving the distribution of 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.

The diagnostic value $d_0(C,H)$ of the absence of cue C with respect to a particular hypothesis H is given by the parallel expression:

$$d_0(C,H) = \frac{f_0(C,H)/f(H)}{\sum_j [f_0(C,H_j)/f(H_j)]}. \tag{9b}$$

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, as is elaborated in the general discussion, the model will not capture configural cue processing effects (Edgell, 1978, 1980). The calculation of diagnostic value is also uninfluenced by the baserate or prior probability of the hypothesis in question, as is the case for the likelihood ratio in Bayes' rule.

4.3. Summation of diagnostic implications of individual cues

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 for a particular hypothesis of the cue pattern taken as a whole. It is assumed that present cue values are given greater weight than are absent cue values in the summation process. This assumption is consistent with previous research investigating judgments of covariation and causation (e.g., Kao & Wasserman, 1993; Schustack & Sternberg, 1981; Shaklee & Mims, 1982; Smedslund, 1963) indicating that the presence of cues or the occurrence of events typically receives more weight in intuitive judgments than does the absence of cues or the non-occurrence of events. The diagnostic value of cue pattern $\mathbf{C}$ for hypothesis H, denoted $d_{\mathbf{C}}(H)$, is given by

$$d_{\mathbf{C}}(H) = (1-\delta) \sum_{i}^{\text{present cues}} d_1(C_i,H) + \delta \sum_{i}^{\text{absent cues}} d_0(C_i,H) \text{ for cues } C_i,$$
$$i = 1,\ldots,N_C, \text{in } \mathbf{C}. \tag{10}$$

The free parameter δ 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. 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.

4.4. Calculation of support

As described above, ESAM's diagnostic value calculation controls for the baserate frequency $f(H)$ of the hypothesis under evaluation, and in this sense is insensitive to overall differences in baserate or prior probability among the competing hypotheses. ESAM accommodates potential baserate 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 $s_{\mathbf{C}}(H) \geqslant 0$, is given by

$$s_{\mathbf{C}}(H) = \left[\alpha\left(\frac{f(H)}{\sum_j f(H_j)} - \frac{1}{N_H}\right) + (1-\alpha)d_{\mathbf{C}}(H)\right]^{\gamma}. \tag{11}$$

The free parameter α 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 baserate (i.e., observed relative frequency in comparison with the alternative hypotheses). The adjustment is positive in the case of high-baserate hypotheses whose relative frequency exceeds $1/N_H$, the value expected under a uniform partition; the adjustment is negative for low-baserate hypotheses. With unequal baserates, α reflects the judge's sensitivity to this consideration. In the special case of equal baserates, this adjustment is zero and the parameter α drops out of the model. The support calculation echoes that of the Bayesian approach, in that it combines considerations of the diagnosticity of the available evidence and the prior probability or baserate of the hypothesis. Note, however, that in contrast to the Bayesian model, the form of ESAM's baserate adjustment is not multiplicative, an assumption that is consistent with research on intuitive use of baserate information (Birnbaum & Mellers, 1983; Novemsky & Kronzon, 1999).

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 baserate, 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 that influences the extremity of the resulting judgments; categorization models often employ a similar parameter (e.g., Nosofsky & Johansen, 2000). 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. The value of this parameter might, for instance, be sensitive to the size of the set of observations upon which the judgments are based and the learning conditions under which those observations were made.

4.5. Enhanced residual discounting

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. As discussed above, Koehler et al. (1997) offered a simple linear discounting model (6) according to which the greater the support for the focal hypothesis, the greater the discounting of support for its alternatives by virtue of their being packed together in the residual. This phenomenon of *enhanced residual discounting* reflects the intuition that the judge is less likely to fully evaluate the extent to which the evidence supports alternative hypotheses when support for the focal hypothesis is high than when it is low. In ESAM, enhanced residual discounting is implemented by restricting the number of cues that are consulted 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 consulted with regard to their contribution to the support for an alternative hypothesis included in the residual.

Put differently, we assume that the probability of a cue value being consulted and contributing to the support for the focal hypothesis is 1 (i.e., each cue is always consulted), but its probability of being consulted and contributing to the support for an alternative hypothesis included in the residual is less than 1. For simplicity, we assume that given a particular level of support for the focal hypothesis, the probability q that a cue will be consulted 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. In terms of expected value, 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:

$$q_{\overline{H}} = \frac{1}{\beta s(H) + 1}. \tag{12}$$

The free parameter β determines how quickly $q_{\overline{H}}$ decreases as $s(H)$ increases.

Support for the residual is then given by

$$s_{\mathbf{C}}(\overline{H}) = \sum_{H_j \text{ in } \overline{H}} \left[\alpha \left(\frac{f(H_j)}{\sum_i f(H_i)} - \frac{1}{N_H} \right) + (1-\alpha) q_{\overline{H}} d_{\mathbf{C}}(H_j) \right]^{\gamma}. \tag{13}$$

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 perceived diagnostic value of the evidence is then adjusted for the baserate of the hypothesis in the usual manner. 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.

Note that in contrast to the linear-discounting model (see (4)–(6) above) of Koehler et al. (1997), in which support for alternatives included in the residual is directly discounted, in ESAM it is only the diagnostic value component of the support calculation that is discounted. This follows from our interpretation of q as the probability of consulting a particular cue in the support assessment process. Hence, in our model, discounting of the residual support term as a whole is non-linear in its relationship to the support for the focal hypothesis. Due to this non-linear relationship, it should be noted that while q can be viewed as an expectation with respect to a stochastic cue-sampling process of the kind described above, the support value resulting from (13) is not the expected support value that such a stochastic process would produce. Despite the difference in form between the current model and that of Koehler et al. (1997), both produce enhanced residual discounting: Support for the residual is discounted more extensively as the support for the focal hypothesis increases.

4.6. Summary of ESAM and comparison to Bayesian approach

To summarize, 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 (baserate adjustment), β (enhanced residual discounting), γ (judgmental extremity), and δ (absent cue weighting). According to the model, cue patterns are assessed one cue at a time. A cue value provides support for a particular hypothesis to the extent that it co-occurs with that hypothesis more frequently than would be expected on the basis of the hypothesis's baserate alone. In this manner, 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 weight placed on absent cues relative to that placed on present cues in the summation process. The free parameter α reflects the extent to which the diagnostic value assessment is adjusted in light of the baserate of the hypothesis under evaluation. The free parameter γ reflects the extremity of the resulting support estimates. The free parameter β reflects the degree to which, as the support for the focal hypothesis increases, the set of cues consulted is restricted in assessing the support for hypotheses included in the residual.

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? While 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 is 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

$$P(H \mid \text{cue pattern } C) = \frac{f(H)\prod_{i=1}^{N_C}(f(C_i,H)/f(H))}{\sum_{j=1}^{N_H}\left[f(H_j)\prod_{i=1}^{N_C}(f(C_i,H_j)/f(H_j))\right]} \text{ for } C_i \text{ in cue pattern } \mathbf{C}, \quad (14)$$

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, and $f(H)$ is the overall frequency with which H was previously observed.

The numerator of (14) 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 baserate 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 (14) described above is that the Bayesian model integrates individual cue values (and considerations of hypothesis baserate or prior probability) in a multiplicative manner, while ESAM uses an additive integration form. A consequence is that in ESAM's additive framework, support tends to increase with the number of cues consulted, while in the Bayesian model it tends to decrease. Furthermore, in the integration process, ESAM accommodates differential weighting of cue absence and cue presence, while in the normative Bayesian approach cue absence and cue presence are logically interchangeable. Because ESAM is not a generalization of the Bayesian model (14) above, 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 in a later section.) 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.

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 above, by contrast, assume conditional independence of cues. The frequentist approach does, however, require a very large sample of previous observations in order 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 above might be more useful in producing reasonably accurate judgments in the face of small sample sizes and limited memory capacity. In the next section, ESAM's fit to new data is assessed and compared to that of alternative models, including those just described.

5. Experimental data

Data from four new experiments were used to assess the proposed model and to compare its performance to that of alternative formulations. In this section we provide an overview of the experiments and their basic results.

5.1. Method

Participants were undergraduate students enrolled in an introductory psychology course at the University of Waterloo, who received course credit in exchange for their participation. The computer-based experiments were conducted in individual sessions taking approximately an hour to complete.

The four experiments shared a common design, procedure, and instructions, which closely followed that of Koehler (2000). Participants played the role of a physician in a simulated medical diagnosis task. Each "patient" presented to participants was known to be suffering from one of three possible flu strains (simply numbered Flu Strain #1, #2, and #3). Participants attempted to diagnose the flu strain of each patient on the basis of six binary symptoms (*cough*, *chills*, *dizziness*, *earache*, *headache*, and *sore throat*) known to be present or absent for that patient.

In the training phase of the experiment, the participant was presented with a series of 300 patients with the task of choosing which of the three flu strains each patient was suffering from. Following their choice, the correct diagnosis was provided. In this manner, participants learned about the probabilistic relationships between the symptoms and flu strains. The training phase instructions warned participants that, as in actual medical practice, it would not be possible to achieve perfect diagnostic accuracy on the basis of the observable patient symptoms. Each participant in a given experiment was presented with an identical set of 300 training trials, but in

different, individually randomized orders. The fixed order in which the six symptoms were listed was also determined randomly for each participant.

In the judgment phase of the experiment that followed, participants were presented with additional patients and, for each, judged the probability that the patient was suffering from a designated flu strain. Probability judgments were made on an 11-point scale ranging from 0 to 100% in increments of 10%, where 100% indicates certainty that the patient has the designated flu strain and 0% indicates certainty that the patient does not have that flu strain. The judgment phase instructions emphasized that the flu strain designated as the target of judgment on a particular trial would be selected arbitrarily, and that its designation should not be taken as having any informational value regarding the patient's diagnosis. No feedback was provided in the judgment phase.

For each of the 64 possible symptom patterns, participants assigned a probability to each of the three possible flu strains, for a total of 192 judgments. The 192 judgments were elicited in an order that was randomly determined for each participant. Specifically, an individual participant's judgments of the three flu strains given a particular symptom pattern were made on separate trials, typically with a large number of intervening judgments. In this design, it is difficult for participants to ensure that, for each symptom pattern, the probability estimates they assign to the three flu strains add consistently to one as required by probability theory. As is discussed by Tversky and Koehler (1994), such a design is suitable for our interest in the perceived support for a designated hypothesis provided by a particular body of evidence; whether participants can revise such initial, essentially independent judgments to meet additivity constraints, when they are made salient, is a separate question not directly addressed in our research. It is worth noting in this regard that while a judge may ensure that any particular set of probability estimates is additive, this does not ensure additivity over different possible partitions of the sample space (Tversky & Koehler, 1994; for examples, see Brenner & Koehler, 1999).

During the training phase, in which outcome feedback is provided on each trial, participants made simple predictive choices rather than probability judgments as they did during the subsequent judgment phase. The training phase is greatly simplified and speeded by asking for a predictive choice rather than an explicit probability judgment on each trial. Koehler (2000) found similar results using either type of response during the training phase of his experiments, with no evidence that the probability judgments are less subadditive when accompanied by outcome feedback.

The four experiments differed only in their cue structure, that is, in the way in which the symptoms serving as the basis of judgment were related to the three possible flu strains. In other words, the experiments differed in the set of training trials presented to participants. Table 1 shows, for each experiment, the frequency with which each flu strain co-occurred with the presence of each of the six symptoms. The number of patients in the training set exhibiting each flu strain is also indicated, from which the frequency of co-occurrence of symptom absence with each flu strain can easily be calculated. In the experiments, the flu strains were simply labeled #1, #2, and #3, but meaningful symptom labels (*cough*, *chills*, *dizziness*, *earache*,

Table 1
Frequency of co-occurrence of flu strain with symptom presence over the 300 training trials of Experiments 1–4

	Flu Strain #1	Flu Strain #2	Flu Strain #3
Experiment 1	($n = 100$)	($n = 100$)	($n = 100$)
Symptom #1	92	29	29
Symptom #2	29	92	29
Symptom #3	29	29	92
Symptom #4	72	39	39
Symptom #5	39	72	39
Symptom #6	39	39	72
Experiment 2	($n = 100$)	($n = 100$)	($n = 100$)
Symptom #1	8	71	71
Symptom #2	71	8	71
Symptom #3	71	71	8
Symptom #4	28	61	61
Symptom #5	61	28	61
Symptom #6	61	61	28
Experiment 3	($n = 100$)	($n = 100$)	($n = 100$)
Symptom #1	92	50	8
Symptom #2	8	92	50
Symptom #3	50	8	92
Symptom #4	72	50	28
Symptom #5	28	72	50
Symptom #6	50	28	72
Experiment 4	($n = 150$)	($n = 100$)	($n = 50$)
Symptom #1	138	29	14
Symptom #2	43	92	14
Symptom #3	43	29	46
Symptom #4	108	39	19
Symptom #5	58	72	19
Symptom #6	58	39	36

headache, and *sore throat*) were substituted for the abstract numerical labels shown in the table. The particular symptom label assigned to the symptom numbers shown in the table was determined randomly for each participant.

In all the experiments, Symptoms #1–3 were more diagnostic than Symptoms #4–6. Within each level of diagnosticity, for each of the three flu strains there was one symptom whose presence implicated that flu strain (except in Experiment 2). Thus, the presence of Symptom #1 implicated Flu Strain #1, Symptom #2 implicated Flu Strain #2, and Symptom #3 implicated Flu Strain #3, each with the same level of diagnosticity. Likewise, the presence of Symptom #4 implicated Flu Strain #1, Symptom #5 implicated Flu Strain #2, and Symptom #6 implicated Flu Strain #3, each with the same lower level of diagnosticity than that afforded by Symptoms #1–3. (In Experiment 2, as elaborated below, cue presence and cue absence were reversed relative to Experiment 1.)

Symptom patterns were determined subject to these constraints on co-occurrence of individual symptoms with the three flu strains. The presence or absence of each of the six symptoms was intended to be conditionally independent of the other symptoms. Generation of a finite set of training trials introduces some minor deviations from perfect conditional independence, but the resulting incidental conditional dependencies among symptoms were weak and unlikely to have been detectable by participants.

The first three experiments all involved equal-baserate flu strains, such that each was observed in 100 of the 300 patients in the training sequence. Experiment 1 served as a baseline study against which results of the remaining studies were compared. As was the case for the diagnostic symptoms in the Koehler (2000) studies, in this experiment the two flu strains not implicated by a particular symptom's presence co-occurred equally often with that symptom; that is, both were equally likely (and less likely than the implicated flu strain) to co-occur with the presence of the symptom. This was true for both the high-diagnosticity and low-diagnosticity set of symptoms.

To examine the influence of people's tendency to focus on cue presence rather than absence, Experiment 2 employed a logically equivalent set of training trials in which symptom presence and absence were reversed relative to that of Experiment 1. For example, in Experiment 1 the presence of Symptom #1 co-occurred with Flu Strain #1 a total of 92 times while its absence co-occurred with Flu Strain #1 a total of 8 times; in Experiment 2, these frequencies were reversed. As a result, in Experiment 2 it was the absence rather than the presence of Symptom #1 that implicated Flu Strain #1, and so on. If people treat symptom presence and absence as equally informative, learning performance during the training phase should be identical in Experiments 1 and 2, and judgments conditioned on symptom absence in Experiment 2 should be identical to those conditioned on symptom presence in Experiment 1. But if people tend to place greater weight on the diagnostic implications of present symptoms, as we suspect, learning may be more difficult in Experiment 2 and the corresponding judgments will not be identical to those found in Experiment 1.

In Experiment 3, in contrast to Experiment 1, alternatives to the flu strain implicated by a particular symptom were not equally supported by that symptom. Instead, symptoms had a graded association with the flu strains, with one flu strain being more likely and another being less likely than the remaining flu strain in the presence of a particular symptom. This design provides some generalizability of any results from Experiment 1, showing that they hold even when the alternatives to the implicated flu strain are not equally likely. Experiment 3 might also impose greater memory demands on the participant than does Experiment 1, because it is no longer sufficient to recall just the extent to which the presence of a symptom implicates one flu strain over the other two, which now also differ from one another in their association with the symptom. To introduce graded symptom associations, the number of times a given symptom co-occurred with the two non-implicated flu strains was redistributed relative to Experiment 1 while its frequency of co-occurrence with the implicated flu strain was maintained.

In contrast to the first three experiments, all of which involved equal baserates, in Experiment 4 unequal baserates were introduced in which Flu Strain #1 occurred

more frequently and Flu Strain #3 occurred less frequently than Flu Strain #2. The training set was constructed by dropping 50 of the original 100 occurrences of Flu Strain #3 in Experiment 1, and adding 50 additional occurrences of Flu Strain #1. The conditional probability of each symptom given each flu strain was equivalent to that in Experiment 1. In addition to providing an opportunity to generalize the basic findings of the earlier experiments to the case of unequal baserates, this experiment also allows examination of the usefulness of ESAM's baserate adjustment parameter α. That is, only in the case of unequal baserates is it meaningful to compare ESAM's fit to the data with that of an otherwise comparable model that does not accommodate adjustment of support for a hypothesis due to its overall baserate of occurrence.

5.2. Results

Initially, the number of participants was 42, 42, 41, and 44 in Experiments 1–4, respectively. As elaborated below, participants who showed poor learning performance or who failed to complete the entire experiment were dropped from our main analyses. The final number of participants was 34, 32, 33, and 35 in Experiments 1–4, respectively.

5.2.1. Learning performance

Our focus is on the probability judgments made following the learning phase of the experiment. Data from the learning phase are used primarily as a means of screening out participants who show signs of having failed to learn the diagnostic relationships between the symptoms and flu strains over the course of the training trials. This is necessary because the current version of ESAM presumes reasonably accurate learning in its use of veridical frequency counts to predict support assessments. The model could, of course, be modified to account for what participants actually learn during the training phase, but for present purposes we restrict our analyses to those participants who exhibit satisfactory learning performance.

One complication that arises in the screening process is that learning performance differed substantially across experiments. As expected, learning performance as measured by proportion of correct diagnoses was generally much lower in Experiment 2, where present and absent symptoms were reversed relative to Experiment 1. This observation is consistent with a tendency to focus on symptom presence. The average proportion of correct diagnoses in the last half of the training trials was 59, 47, 56, and 60% in Experiments 1–4, respectively. By comparison, at asymptote, a Bayesian analysis assuming conditional independence of cues (14) would be expected to achieve an accuracy rate of 74, 74, 88, and 79% in Experiments 1–4 respectively. In short, the observed accuracy during the learning phase of the experiments was well above chance but well below the theoretical maximum expected at the end of the training sequence, perhaps due to the complex nature of the task (involving 6 cues and 3 possible outcomes).

We wished to eliminate from further analyses any participants who exhibited poor learning performance, but due to the generally lower proportion of correct

diagnoses, adopting a fixed cutoff value would result in the exclusion of a larger proportion of participants in Experiment 2, arguably producing a more highly selected sample not comparable to that included in the other experiments. To avoid this problem, we chose to exclude the same proportion of relatively poor-performing participants from each experiment on the basis of their proportion of correct diagnoses over the last half of the training trials. By this measure, approximately 20% of participants in Experiment 2 had accuracy scores near or below that expected by chance alone (i.e., proportion correct = 1/3). By dropping these participants, the resulting sample included only participants with a greater proportion of correct diagnoses than that expected by chance. To maintain comparability of samples, the same proportion of participants was dropped from Experiments 1, 3, and 4, that is, the lowest-ranking 20% of participants as measured by proportion of correct diagnoses over the last half of the training trials.

5.2.2. Probability judgments

Two additional participants in Experiment 2 were dropped because their judgment data were incomplete, presumably due to their quitting before the end of the experiment. We fit ESAM and alternative models to the set of mean probability judgments in each experiment, and also to each individual's judgment data as a way of conducting inferential statistical tests on parameter values and fit indices.

For each experiment, the set of 192 mean probability judgments (the mean probability assigned to each of the three flu strains given each of the 64 possible symptom patterns) was computed. To assess the general accuracy of these judgments, the corresponding Bayesian values were calculated as described above (14), assuming conditional independence among the symptoms. Fig. 1 (unfilled circles) plots the mean probability judgments against the Bayesian values separately for each experiment. The accuracy of the mean judgments is generally quite impressive, though of course the accuracy of individual judgments will tend to be less so. The correlation between the mean judgments and the Bayesian values is .95, .85, .90, and .93 in Experiments 1–4, respectively. As would be expected given the imperfect correlations, judgments associated with high Bayesian values tended to be too low, while those associated with low Bayesian values tended to be too high. What is perhaps less obvious is that the judgments generally tended to be too high, as indicated by a larger proportion of the points on the scatterplot falling above rather than below the identity line. This pattern is to be expected to the extent that such judgments exhibit systematic subadditivity.

In Fig. 2, the degree of subadditivity is measured by the total probability T assigned to the three possible flu strains given a particular symptom pattern. The figure shows the mean value of T (filled circles) as a function of the number of present symptoms in the symptom pattern. The horizontal line designates the value $T = 100\%$ expected under the Bayesian analysis. Classification-learning models that normalize the output associated with each category to arrive at a probability judgment, such as ALCOVE (see Nosofsky et al., 1992) or Gluck and Bower's (1988) adaptive network model, also yield the prediction $T = 100\%$. Consistent with support theory, however, the figure shows that participants' judgments were generally

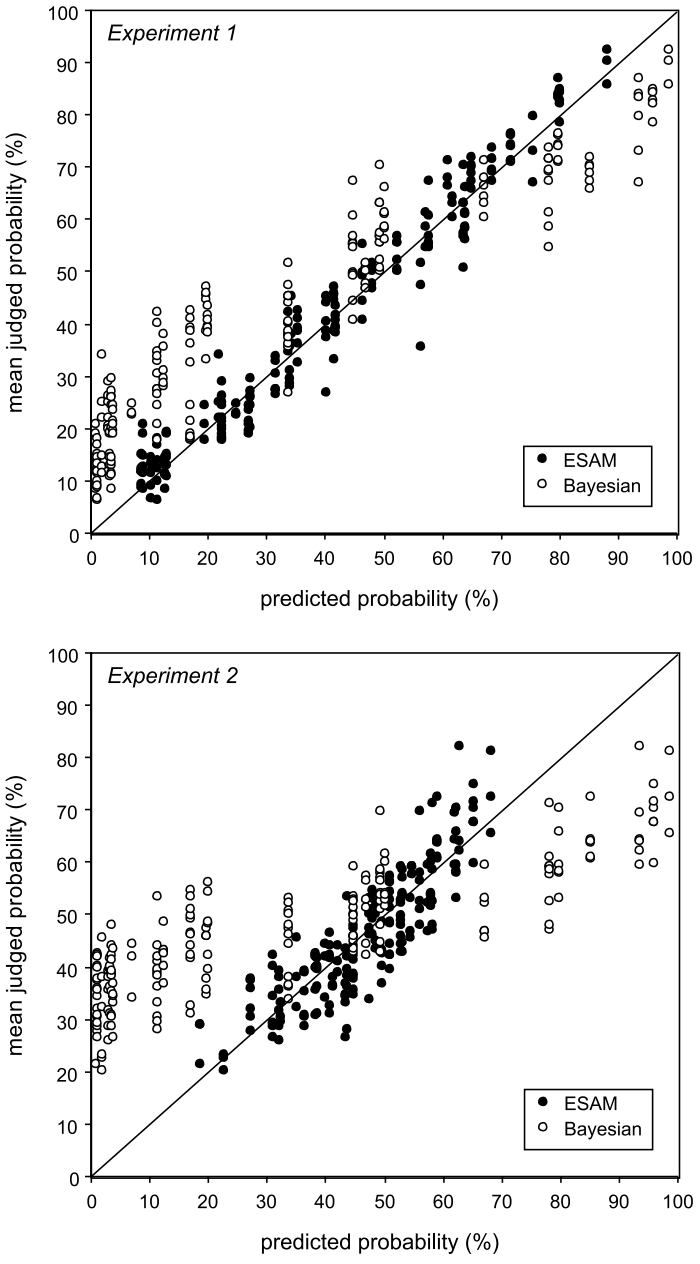

Fig. 1. Mean probability judgments in Experiments 1–4 versus ESAM and Bayesian predictions.

subadditive as indicated by $T > 100\%$. Furthermore, consistent with the results of Koehler (2000), the value of T increases systematically with the number of present symptoms in the symptom pattern upon which the judgments are based. This

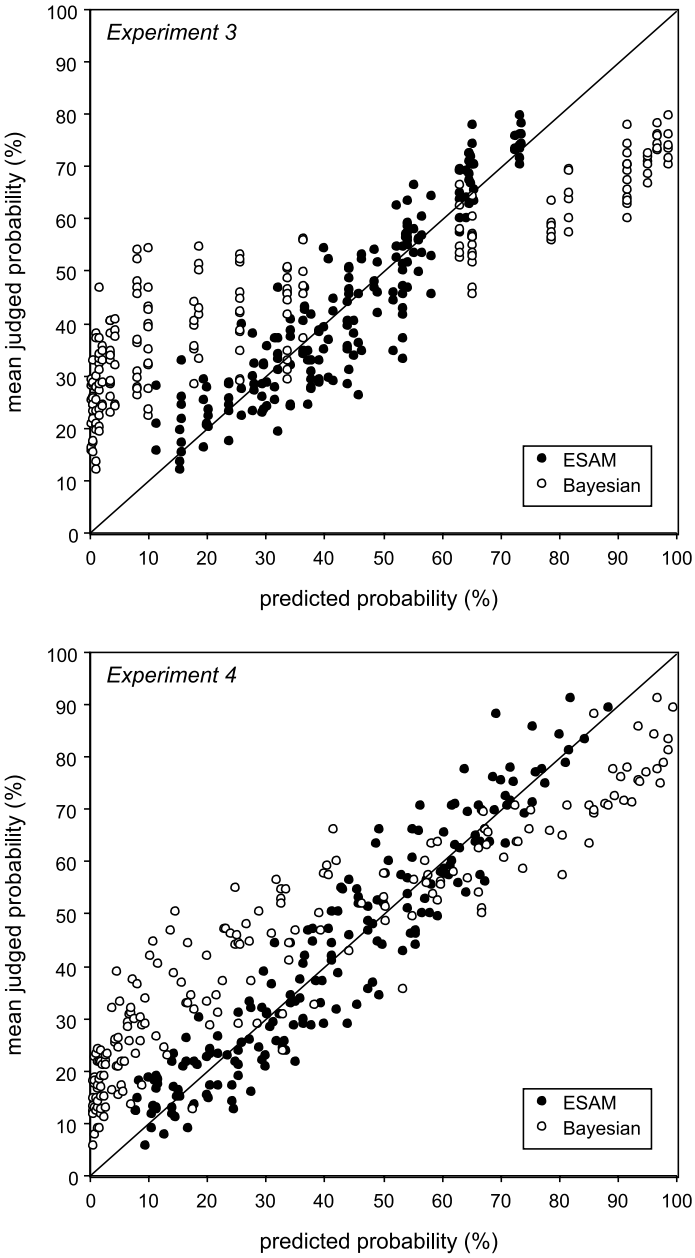

Fig. 1. (*continued*)

observation is one of the key findings that ESAM is designed to accommodate. We are not aware of any current models of classification learning that, without further modification, are able to account for variance in T.

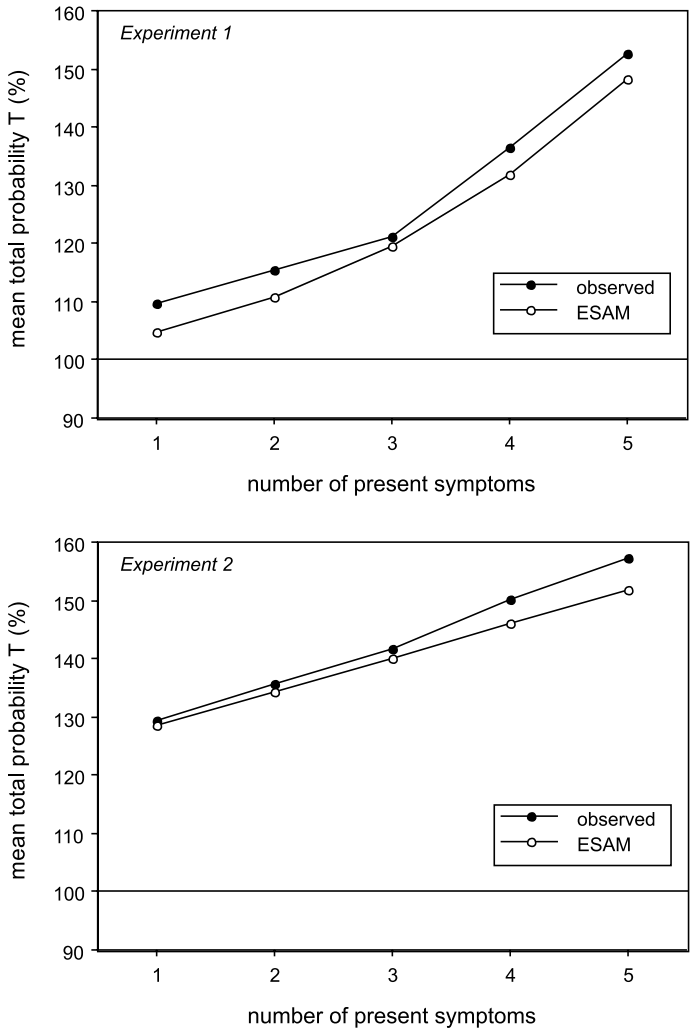

Fig. 2. Predicted and observed mean values of T (the total probability assigned to the three possible flu strains) as a function of the number of present symptoms in the symptom pattern in Experiments 1–4.

Fig. 2 shows that, in each experiment, T increases in an approximately linear fashion as the number of present symptoms increases from 1 to 5. We have excluded from the figure points corresponding to the special cases of 0 and 6 present symptoms, that is, the two cue patterns in which all the symptoms are either absent or present. The value of T tends to be less predictable for these special cases, in part because the means are based on fewer observations, and in part because participants appear to treat them somewhat differently than the other symptom patterns. In particular, T tends to be substantially lower in the case in which all possible symptoms are present than would be expected based on the general increasing trend. Koehler

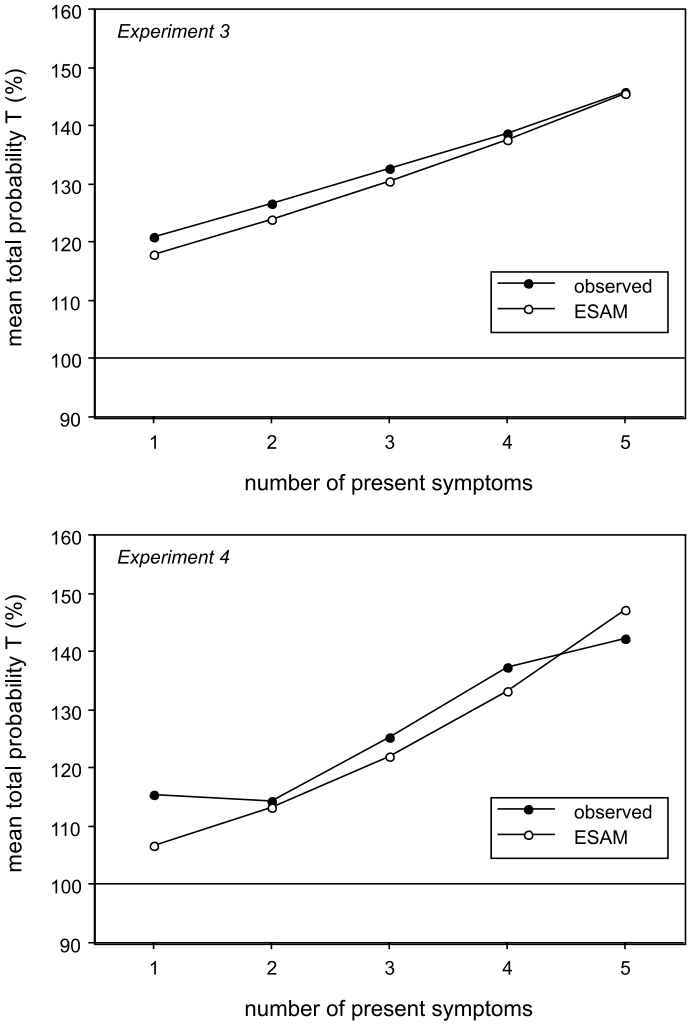

Fig. 2. (*continued*)

(2000) observed the same phenomenon and suggested that these patterns may be perceived as especially uninformative with regard to the patient's diagnosis and hence as not very supportive of any of the flu strains. It is worth noting that without further modification that takes these special cases into account, ESAM's general assumption of support accumulation will fail to capture this aspect of the data. Of course, these special cases represent only 2 of the 64 possible symptom patterns evaluated in each experiment.

Recall that in each experiment, there were three high-diagnosticity and three low-diagnosticity symptoms. While T was generally expected to increase with the number of present symptoms, the assumptions upon which ESAM is based imply that this

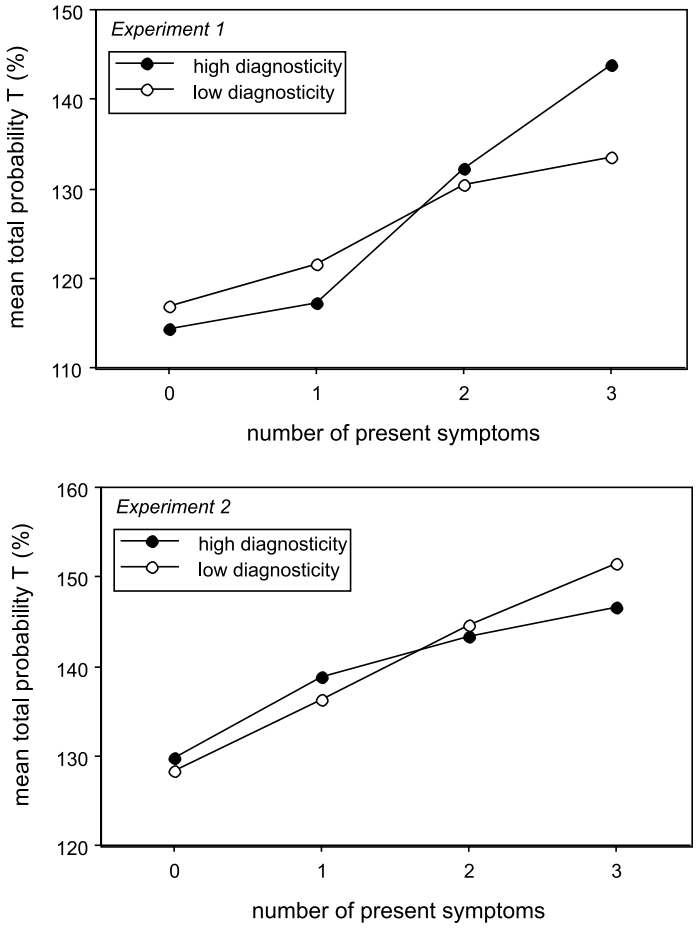

Fig. 3. Mean value of T as a function of the number of present high- and low-diagnosticity symptoms.

pattern should be more pronounced in the case of high-diagnosticity symptoms than in the case of low-diagnosticity symptoms. If a symptom's diagnostic value determines how much its presence contributes to the support for a hypothesis, then the presence of high-diagnosticity symptoms will generally produce greater support than the presence of corresponding low-diagnosticity symptoms, assuming that symptom presence is weighted more heavily than absence. Because alternatives in the residual are discounted more heavily as the support for the focal hypothesis increases, this leads to the prediction that T should increase more quickly as a function of the number of high-diagnosticity symptoms present in the symptom pattern than as a function of the number of low-diagnosticity symptoms present. Fig. 3 shows that this is indeed the case in all experiments except Experiment 2, where the much greater difficulty participants had in the training trials might have made it more difficult to detect the difference in diagnosticity between the two sets of symptoms. In each of the

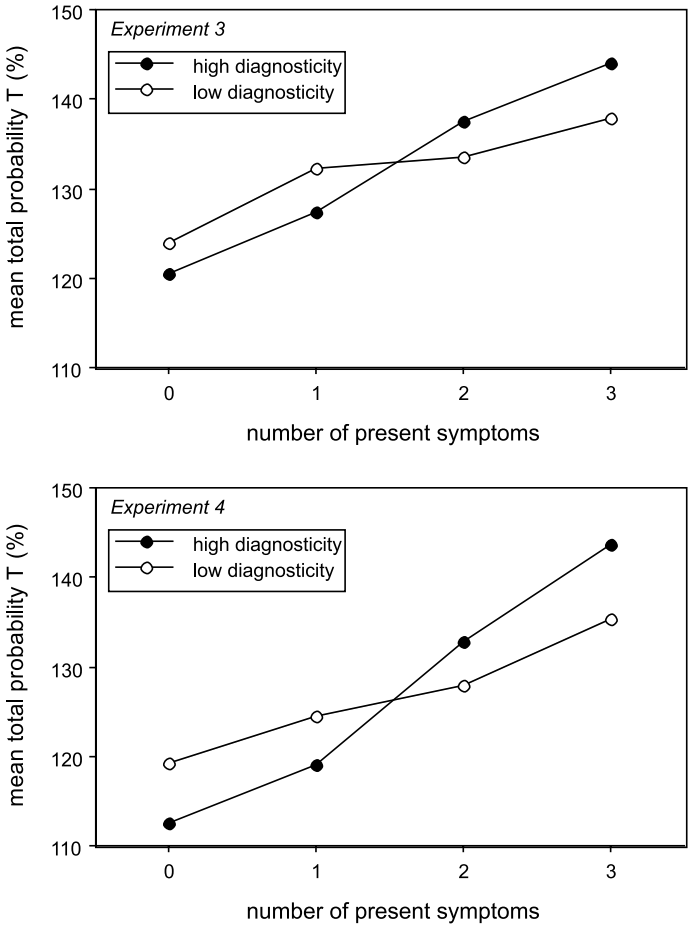

Fig. 3. (*continued*)

remaining experiments, the correlation between the number of high-diagnosticity symptoms present and T is significantly greater than that between the number of low-diagnosticity symptoms and T.

6. Model fitting

In this section we first assess how well ESAM is able to fit the data from the four experiments, and then compare its performance with that of alternative models.

6.1. Fitting ESAM to mean judgments

We first fit ESAM to the set of mean probability judgments from each experiment. Best-fitting parameter values were estimated simultaneously using an iterated search

algorithm that minimized the squared difference between predicted and observed probability judgments. The top section of Table 2a shows the estimated parameter values and provides a measure of the goodness of ESAM's fit to the mean data in each experiment. The observed mean judgments are plotted against the resulting predicted values in Fig. 1 (filled circles). Both the table and the figure indicate that ESAM was able to quite closely reproduce the 192 observed mean judgments in each experiment.

As a way of interpreting the measure of ESAM's fit to the data as listed in Table 2a, consider the following. Because ESAM's predictions are based on the veridical frequency counts determined by the cue structure in a given experiment, and because the cue structures used in Experiments 1 and 2 (and, to a lesser extent, in Experiment 3) are highly symmetric, it turns out that ESAM provides only 40 (64 in Experiment 3) unique predicted judgments which are repeated over the full set of 192 possible judgments in each experiment. For example, the predicted probability

Table 2a
Estimated parameter values and index of fit of 4-parameter ESAM and alternative models to mean judgment data from Experiments 1–4

Model	Experiment	α	β	γ	δ	RMSE	t
ESAM	1	—	0.10	2.96	0.22	4.82	
	2	—	0.66	1.14	0.27	6.19	
	3	—	0.32	1.62	0.27	6.85	
	4	0.25	0.28	2.52	0.22	6.88	
Pattern similarity	1	—	0.38	0.96	0.27	6.29	5.35***
	2	—	1.00	0.56	0.42	6.57	2.98***
	3	—	0.74	0.64	0.34	6.93	0.72
	4	0.00	0.26	0.98	0.36	10.51	5.17***
Redundant baserate	4	0.00	0.80	0.69	0.00	16.4	8.67***
Uncorrected diagnosticity	4	−0.06	0.82	0.61	0.00	16.5	8.87***
Constant adjustment	4	−0.22	0.16	1.95	0.23	8.20	3.45***
Constant discounting	1	—	0.88	3.58	0.22	5.41	2.63***
	2	—	0.37	1.47	0.37	6.67	2.82***
	3	—	0.78	2.02	0.27	7.20	1.82*
	4	0.24	0.87	3.08	0.22	7.19	0.56
Bayesian ESAM	1	—	2.03	0.38	0.61	5.80	0.95
	2	—	3.54	0.15	0.25	5.81	−1.69*
	3	—	2.37	0.20	0.42	6.48	−0.69
	4	0.82	0.51	0.34	0.62	6.53	0.51

*$p < .10$; **$p < .05$; ***$p < .01$. *Note*. α, baserate adjustment; β, enhanced discounting; γ, extremity; δ, absent cue weighting; RMSE, root mean squared error of ESAM's fit to the set of 192 mean probability judgments (on the 0–100% judgment scale) in each experiment; t, paired t test comparing fit of alternative model to that of ESAM with equal number of free parameters, based on individual judgment data.

assigned to Flu Strain #1 given a cue pattern in which only Symptom #1 is present is identical to that assigned to Flu Strain #2 given a cue pattern in which only Symptom #2 is present, and so on. (The same holds for predictions from the Bayesian approach.) Thus one reasonable baseline against which to compare ESAM's performance is the fit exhibited by a model that perfectly reproduces the cell means for each of these 40 (or 64 in Experiment 3) cases. The predictive error associated with this model is attributable to variance within a given cell, for example, if the mean probability assigned to Flu Strain #1 given a cue pattern in which only Symptom #1 is present does not exactly coincide with that assigned to Flu Strain #2 given a cue pattern in which only Symptom #2 is present, and so on. The RMSE associated with this cell-means model is 3.41, 5.02, and 4.64 in Experiments 1, 2, and 3, respectively. (Due to the unequal hypothesis baserates, ESAM is capable of producing 192 unique predicted judgments in Experiment 4 and hence the corresponding cell-means model would perfectly fit the data.) By comparison, ESAM produces a fit that is not that much worse using only three free parameters. This analysis also highlights that a poorer fit would be expected in Experiments 2 and 3 than in Experiment 1 solely on the basis of the greater within-cell variance in the former than in the latter.

Fig. 1 indicates that ESAM provides a substantially better fit to the data than do the Bayesian values, which serve as a benchmark of how predictable people's judgments are from the normative values computed from the same set of frequentistic observations serving as input for ESAM. Because they rely on the same input, the Bayesian and ESAM predicted probabilities are highly correlated with one another as well as with the observed judgments. The partial correlation between ESAM's predictions and the observed mean judgments, controlling for the Bayesian values, is .80, .49, .59, and .62 in Experiments 1–4, respectively, indicating that ESAM is able to account for substantial variance in the observed judgments that is not captured by the Bayesian model.

A critical advantage of ESAM over the Bayesian model, of course, is that ESAM is intended to account for systematic variance in the subadditivity of elementary judgments as measured by T, which according to the Bayesian approach should always equal 1 (or 100%). Fig. 2 shows that ESAM nicely reproduces the tendency for T to increase with the number of present symptoms in the symptom pattern. The values of T predicted by ESAM in this figure are computed from its predictions regarding the individual judgments, rather than on the basis of direct fitting of the model to the observed T values themselves. The good fit achieved by this approach provides some corroborating evidence for the postulated principles of support assessment on which ESAM is based. Specifically, ESAM's assumption that greater weight is placed on cue presence than on cue absence produces generally greater support values for the focal hypothesis as the number of present cues in the cue pattern increases, and its assumption of enhanced residual discounting then leads to greater discounting of support for alternatives included in the residual (and hence greater values of T) as the support for the focal hypothesis increases with the number of present cues.

Inspection of Table 2a (or Fig. 1) shows that ESAM achieved a better fit to the judgments in Experiment 1 than in any of the other experiments. This result may

be related to the observation that learning performance and judgment accuracy tended to be somewhat higher in this experiment. Experiment 1 arguably imposes lower memory demands on participants because it involves a simpler cue structure and equal baserates. Because our implementation of ESAM assumes perfect accuracy in encoding and retrieval of observed frequencies, the model will tend to produce a better fit to judgments under conditions that more closely meet this assumption, as appears to have been the case in Experiment 1.

The estimated parameter values for ESAM in Table 2a are generally consistent with two expectations based on the support assessment principles introduced earlier. First, the observation of $\beta > 0$ indicates that: (a) support for hypotheses included in the residual is discounted; and that (b) the degree of discounting increases as the support for the focal hypothesis increases. Second, the observation of $\delta < 1/2$ indicates that absent symptoms are given less weight than present symptoms in the support assessment process.

Further assumptions of ESAM are also corroborated by the estimated parameter values. The observation of $\alpha > 0$ in Experiment 4 (the only experiment in which baserate varied) is consistent with some degree of baserate sensitivity, such that the support for a hypothesis based on the diagnostic value of the available cues is adjusted in light of the hypothesis's baserate. The observation of $\gamma > 1$ suggests that the final computation of support associated with a hypothesis tends to be more extreme than the initial calculation of diagnosticity adjusted in light of baserate specified by ESAM.

6.2. Fitting ESAM to individual judgments

To further investigate these parameter values, and in particular how they vary across the four experiments, we fit ESAM to each individual participant's set of 192 probability judgments. Obviously, these judgments are much noisier than the aggregate set of mean judgments in each experiment, and as a result the fit of ESAM is not nearly as good. The individual parameter estimates, however, tended to be reasonably stable across participants in an experiment. Table 3 shows the median estimate for each parameter value in a given experiment and provides the median goodness of fit of ESAM to the individual data.

Because α drops out of ESAM in the case of equal baserates (as in Experiments 1–3), we cannot examine how this parameter varies across experiments. We did examine the relationship between a participant's learning performance (over the second half of the training trials) and his or her estimated α value in Experiment 4, suspecting that participants who exhibited better learning performance would also exhibit greater baserate sensitivity. The correlation between learning performance and α, however, was not statistically different from zero and in fact fell in the direction opposite to our expectations, $r = -.14$.

Individual estimates of β tended to be negatively correlated with estimates of γ in all four experiments. This is sensible because judgments tend to become more extreme as β increases, which can be compensated for by decreasing γ, which has the effect of making the judgments less extreme. Although we had no strong expectations

on this point, it turned out that β varied significantly across experiments, $F(3, 130) = 5.19$, $p < .01$. Specifically, the mean value of β was significantly greater in Experiment 2 than it was in Experiments 1, 3, or 4. However, this is partly due to the strong relationship between β and γ. When γ is used as a co-variate, β varied only marginally between experiments, $F(3, 129) = 2.25$, $p < .10$.

The greater the value of β, the greater is the extent of discounting of support for hypotheses in the residual, and hence the greater the value of T. Given this observation, it is not surprising to find that the estimated value of β for an individual correlated with his or her mean value of T, even after partialling out the effects of experiment and controlling for the value of the γ parameter, though the correlation is fairly weak, $r = .20$. However, the fit to several of the participants' data yielded extreme values of the β parameter. Nine participants had standardized scores above 2.50 on the β parameter, and so could be considered outliers. After removing these participants from the analysis, the relationship is much more strongly pronounced, with the partial correlation being $r = .70$. Given this, it seems likely that the differences in β across experiments is attributable to the greater degree of subadditivity in Experiment 2 than in Experiments 1, 3, and 4, though of course these differences in subadditivity still need to be explained.

The value of γ determines the extremity of the support values and hence the resulting probability judgments, and can be taken as a measure of confidence in one's judgmental accuracy. Because we anticipated that participants would generally find the diagnostic task more difficult in Experiments 2 and 3 than in Experiment 1 (as discussed above), we expected that γ would be lower in the former than in the latter. The value of γ did vary significantly across experiments, $F(3, 130) = 11.9$, $p < .001$, with γ being significantly lower in Experiments 2 and 3 than in Experiments 1 and 4 (using β as a co-variate did not affect this result). Furthermore, within each experiment, γ was the only parameter whose value correlated with the participant's accuracy in the learning phase of the experiment ($r = .53$, .83, .61, and .55 in Experiments 1–4, respectively). Perhaps not surprisingly, those participants who exhibited better learning performance on the training trials subsequently made more extreme probability judgments.

Our main prediction regarding δ was that it would tend to be less than 1/2, indicating greater weight being placed on present than on absent symptoms. We speculated that the value of δ might tend to be greater in Experiment 2 than in the other experiments because in that experiment it was symptom absence that picked out a particular flu strain as being more likely than its alternatives. In fact, the value of δ did not differ significantly across experiments, $F(3, 133) = 1.87$, n.s., although both the median and mean value of δ was greater in Experiment 2 than in any of the other experiments. The contrast comparing the value of δ in Experiments 1 and 2 (in which everything else was held constant except the reversal of absent and present symptoms) did show Experiment 2 to yield significantly higher values of δ, $t(64) = 2.37$, $p < .05$. We also suspected that those participants in Experiment 2 who paid greater attention to absent symptoms might exhibit better learning performance. Evaluation of this relationship is complicated by the presence of 3 individuals with extreme δ values (standardized scores greater than 2). When these outliers are

included, the correlation between δ and learning performance is non-significant, $r = .24$, but when they are excluded the correlation is much stronger, $r = .56$, $p < .01$.

6.3. Comparing ESAM to alternative models

In this section we consider alternative models that differ in key assumptions from ESAM, and assess their fit to the data. Wherever possible, we attempt to formulate models that differ by only one key assumption from and which have the same number of free parameters as ESAM. In such cases, the assumption being tested receives some corroboration if ESAM outperforms the comparison model in its fit to the data. Parameter estimates and indices of fit to the mean judgment data for ESAM and each of the 4-parameter alternative models considered in this section are presented in Table 2a. Inferential statistical tests for each experiment comparing the fit of these alternative models to that of ESAM, based on fitting the models to the individual participant data, are also reported in Table 2a. Some of the alternative models we consider require only three free parameters, in which case we compare their performance to a simpler 3-parameter version of ESAM in which absent cues are ignored. Results of fitting the 3-parameter models to the judgment data are reported in Table 2b.

6.3.1. Underweighting cue absence

To evaluate ESAM's assumption that absent cues receive less weight in the judgment than do present cues, we fit two simpler models to the data, one that completely ignores absent cues (i.e., $\delta = 0$), and one that places equal weight on cue presence and absence (i.e., $\delta = 1/2$). As would be expected given that it has an extra free parameter, ESAM outperforms both of these simpler models in terms of fit to the data. As shown in Table 2b, the *absent cue neglect model* and the *equal weighting model* fit the data about equally well on the whole; each outperforms the other in some experiments, and the fit to the individual versus the mean data from the same experiment produces some contradictory results. A comparison of the fit of the two models to the data as a whole (with experiment as a covariate) reveals no significant difference, $F(1, 133) < 1$. It should be noted that the equal weighting model cannot reproduce the tendency for T to increase with the number of present symptoms as depicted in Fig. 2. The absent cue neglect model is a reasonable candidate for a simpler 3-parameter version of ESAM that retains the ability to reproduce this effect. When we consider other alternative models that require only three free parameters (Table 2b), we will use this special case of ESAM in which absent cues are ignored as the basis for performance comparisons.

6.3.2. Evidence decomposition

ESAM assumes the support for a hypothesis conveyed by a pattern of cues is assessed on a cue-by-cue basis rather than in terms of the cue pattern as a whole. To evaluate this assumption, we constructed two models, in the frequentist vein as discussed earlier, that maintain frequency counts for entire cue patterns instead of

Table 2b
Estimated parameter values and index of fit of 3-parameter ESAM (absent cue neglect) and alternative models to mean judgment data from Experiments 1–4

Model	Experiment	α	β	γ	δ	RMSE	t
ESAM (absent cue neglect)	1	—	0.09	2.63	—	5.91	
	2	—	0.83	0.94	—	7.09	
	3	—	0.33	1.43	—	7.65	
	4	0.25	0.25	2.20	—	8.64	
Equal cue weighting	1	—	0.12	3.09	—	6.48	−0.06
	2	—	0.65	1.14	—	6.74	−1.65
	3	—	0.32	1.64	—	7.55	−0.06
	4	0.25	0.32	2.65	—	8.04	0.56
Pattern matching	1	—	1.06	0.50	—	16.5	6.03***
	2	—	1.38	0.23	—	27.4	11.0***
	3	—	1.15	0.23	—	29.1	11.7***
	4	0.34	2.19	0.43	—	19.4	10.3***
Uncorrected diagnosticity	4	—	0.80	0.69	0.00	16.5	7.63***
No baserate adjustment	4	—	0.14	2.45	0.21	8.28	1.72*

*$p < .10$; **$p < .05$; ***$p < .01$. *Note.* α, baserate adjustment; β, enhanced discounting; γ, extremity; δ, absent cue weighting; RMSE, root mean squared error of ESAM's fit to the set of 192 mean probability judgments (on the 0–100% judgment scale) in each experiment; t, paired t test comparing fit of alternative model to that of ESAM with equal number of free parameters, based on individual judgment data.

separate counts for each cue. The first is a simple *pattern matching model*, in which the diagnostic implications of a target cue pattern are evaluated on the basis of how many times that exact pattern was previously observed in conjunction with each outcome (hypothesis), using the same calculation (9a) and (9b) as that used by ESAM, only now as applied to the frequency counts for entire cue patterns rather than for an individual cue. Because it evaluates cue patterns in their entirety, the pattern matching model has no equivalent of ESAM's δ parameter; consequently, its performance should be compared to the 3-parameter version of ESAM (absent cue neglect model). As shown in Table 2b, the pattern matching model does not fit the data nearly as well as the comparable 3-parameter ESAM model that assumes neglect of absent cues, $F(1, 133) = 242$, $p < .001$. Obviously, it also performs less well than the full 4-parameter ESAM model.

The pattern matching model relies only on previous observations with an exact match to the target cue pattern, and as a result ignores all but a small proportion of the previous observations. We also constructed a 4-parameter *pattern similarity model* (cf. Medin & Schaffer, 1978) in which observations of cue patterns similar to the target cue pattern serving as evidence also contribute to the support it provides. In this model, the similarity of the target cue pattern to each possible pattern (for which frequency observations are stored in memory) is computed as δ^m where m is the number of "mismatches" between cue values in the cue patterns being

compared. The diagnostic implication of each possible cue pattern is calculated as in (9a) and (9b) using frequency counts for entire cue pattern, then weighted by its similarity δ^m to the target cue pattern serving as the basis of the judgment. The similarity-weighted sum across all possible cue patterns yields $d_{\mathbf{C}}(H)$, which is then translated into support in the same manner (11) as in ESAM. Note that the simple pattern matching model described above is a special case of the pattern similarity model with $\delta = 0$. The pattern similarity model did not generally fit the data as well as ESAM, $F(1, 133) = 28.6$, $p < .001$. Table 2a shows that ESAM's advantage over the pattern similarity model is substantial in Experiments 1 and 4, less pronounced in Experiment 2, and negligible in Experiment 3.

6.3.3. Baserate-corrected diagnosticity calculation

ESAM's calculation (9a) and (9b) of the diagnostic value of a cue "corrects" for overall baserate differences among the hypotheses under consideration, such that the observation that a cue co-occurs frequently with a particular hypothesis is discounted as an indicator of diagnosticity to the extent that the hypothesis has a high baserate or prior probability. To evaluate ESAM's assumption that perceptions of diagnostic value are in this sense corrected in light of baserate differences among hypotheses, we constructed an *uncorrected diagnosticity model* that simply normalizes the co-occurrence frequencies without dividing each by the corresponding hypothesis baserate. For example, in the case of present cues we have

$$d_1(C,H) = \frac{f_1(C,H)}{\sum_j f_1(C,H_j)}. \tag{15}$$

This model incorporates baserate differences directly in the diagnostic value calculation (i.e., does not remove baserate influences from the calculation) and as such does not require a separate baserate adjustment parameter α. The uncorrected diagnosticity model is equivalent to ESAM in the case of equal baserates. Table 2b shows that the uncorrected diagnosticity model does not fit the data from Experiment 4 (the only experiment with unequal baserates) nearly as well as the 3-parameter absent cue neglect version of ESAM. For a more direct comparison to the 4-parameter ESAM model, we constructed a *redundant baserate model* that maintains the additive baserate adjustment associated with the α parameter and as a result is potentially influenced twice by the hypothesis baserate (once in the diagnosticity calculation and once in the additive adjustment). We also constructed a 4-parameter *uncorrected diagnosticity model* in which α is not multiplied by the relevant hypothesis baserate but instead serves as a constant additive adjustment (as in the constant adjustment model discussed below). As is shown in Table 2a, both of these alternative models fit the data from Experiment 4 less well than does ESAM.

6.3.4. Normalized diagnosticity calculation

ESAM's calculation (9a) and (9b) of diagnostic value also involves normalization over the set of possible hypotheses. This normalization ensures that each cue in the cue pattern is weighted equally in terms of its contribution to the support of a hypothesis despite differences in overall cue frequency, that is, how prevalent the cue's

presence (or absence) is across the different hypotheses. We evaluated this assumption by constructing an *unnormalized diagnosticity model*. In the case of present cue values, for example, diagnostic value is given by

$$d_1(C,H) = f_1(C,H)/f(H). \tag{16}$$

The unnormalized diagnosticity model produces different predictions than ESAM only when the cues under consideration vary in their overall frequency or prevalence, which was not the case in Experiments 1–4. Cues did vary in frequency, however, in Koehler's (2000) Experiment 3, which provides an opportunity to evaluate the merits of normalization. As shown in Table 4, ESAM provides a better fit to the data from this experiment than does the unnormalized diagnosticity model, supporting the normalization assumption, $t(15) = 3.31$, $p < .01$. This result suggests that assessment of the diagnostic implications of a cue is corrected for the cue's overall prevalence.

6.3.5. Baserate adjustment of support

ESAM accommodates baserate sensitivity in the support assessment process via the additive adjustment associated with the α parameter. We evaluate this assumption by considering alternative models that make no adjustment for baserate. We first constructed a *constant adjustment model* in which the diagnosticity calculation is subject to a constant additive adjustment independent of baserates:

$$s_{\mathbf{C}}(H) = [\alpha + (1-\alpha)d_{\mathbf{C}}(H)]^{\gamma}. \tag{17}$$

As shown in Table 2a, ESAM's baserate adjustment assumption yields a better fit to the data from Experiment 4 (the only one with unequal baserates) than does the constant-adjustment model. As it happens, the best fit for the constant adjustment model for the Experiment 4 data is achieved with $\alpha = 0$, such that the resulting model is equivalent to a 3-parameter no-adjustment model for this data set. Thus we also list a *no baserate adjustment model* that sets $\alpha = 0$ among the 3-parameter models in Table 2b. This model performs marginally worse than the 3-parameter version of ESAM that neglects absent cues, reflecting the fact that both of these models exhibit approximately equivalent drops in performance relative to the full 4-parameter version of ESAM. Despite the relatively weak corroboration received from the present results, we suggest that it is worth retaining ESAM's ability to accommodate baserate sensitivity because judgments of probability have been observed to be sensitive to baserates under at least some conditions (see Novemsky & Kronzon, 1999).

6.3.6. Enhanced residual discounting

ESAM assumes that discounting of support for alternatives in the residual (via restriction of the set of cues consulted in the assessment process) increases with the support for the focal hypothesis. The general assumption of enhanced residual discounting is already well supported by previous research (Brenner & Koehler, 1999; Koehler et al., 1997), but in the present treatment we have instantiated this assumption in a different form that merits further testing. To test the enhanced discounting assumption as implemented in ESAM, we constructed a *constant*

discounting model in which the probability $q_{\overline{H}}$ of a cue being consulted in the calculation of residual support (13) does not depend on the support for the focal hypothesis (as in Eq. (12)), but instead is set to a constant $q_{\overline{H}} = \beta$. Table 2a shows that the constant discounting model does not fit the data as well as ESAM, $F(1, 133) = 11.8$, $p < .001$.

6.3.7. Perfect memory

A major, arguably unrealistic, simplifying assumption of ESAM is its reliance on a perfectly accurate set of stored frequency counts of cue and hypothesis co-occurrence, that is, an assumption of perfectly reliable encoding and retrieval of information acquired from experience (i.e., training). To evaluate the costs of this assumption in terms of fit to the observed data, we constructed a model that relies on an imperfect representation of stored frequency counts, which can be thought of either as due to unreliability of encoding during training or inaccuracy at the time of later retrieval. By comparing ESAM's fit to the data when provided with inaccurate information as input (the *imperfect memory model*) to the fit achieved by the original version of ESAM, we can measure the loss of fit introduced by our simplifying assumption of perfect memory.

On the assumption of perfectly reliable encoding and retrieval, the observation of a particular cue value in conjunction with a given hypothesis (or outcome) adds 1 to the frequency count associated with their co-occurrence and 0 to all other co-occurrence frequencies for which the model maintains counts. Imperfect memory could introduce two types of errors in the frequency counts, depending on whether the model "forgets" which cue value or which hypothesis (outcome) was observed. Let a represent the probability of correctly encoding and retrieving the observed hypothesis or outcome, and assume that the probability of one of the other hypotheses being incorrectly encoded or retrieved instead is equal for all the remaining hypotheses; that is, each alternative to the observed hypothesis has a probability of $(1 - a)/(N_H - 1)$ of being incorrectly associated with the observed cue value in the stored frequency counts. Likewise, let b represent the probability of correctly encoding and retrieving the observed cue value (i.e., as present or absent) in association with the observed hypothesis (or outcome); therefore the probability of incorrectly encoding or retrieving the cue value (i.e., mistaking presence for absence or vice versa) is $1 - b$. Assuming these two types of errors occur independently, calculating their joint probabilities is straightforward, but to maintain a tractable number of free parameters we estimate a assuming $b = 1$ and vice versa. As long as a and b are relatively close to 1 when estimated separately, this simplifying assumption is unlikely to have a large effect on the results.

The first imperfect memory model, then, involves errors in encoding or retrieval of the appropriate hypothesis, where a represents the probability that the observed hypothesis is properly counted. The original version of ESAM assuming perfect memory sets $a = 1$. The imperfect memory model allowing inaccurate hypothesis counts is bound to provide a better fit to the data because it includes a as an additional free parameter. The question is whether the improvement in fit justifies inclusion of the additional free parameter. Table 2c shows the fit of this 5-parameter

imperfect memory model along with the fit of the original 4-parameter version of ESAM for comparison. The imperfect memory model fits the data no better than the original version of ESAM in Experiment 1, marginally better in Experiment 2, and significantly better in Experiments 3 and 4. As expected, its fit to the overall data is significantly better than that of ESAM assuming perfect memory, $F(1, 133) = 12.0$, $p < .001$. The relatively modest magnitude of improvement in the fit to the mean judgment data shown in Table 2c, however, suggests that the cost of ESAM's perfect memory assumption may be outweighed by the benefits of its relative simplicity.

To evaluate further our interpretation of the a parameter as reflecting memory errors in encoding or retrieval of the training trial information, we examined whether the estimated value of an individual's a parameter was associated with the individual's choice accuracy in the training phase of the experiment. After all, if the a parameter measures the fidelity of an individual's memory, then it should correlate with his or her training accuracy. As explained before, γ is also correlated with training accuracy. After controlling for γ and experiment, the correlation between a and training accuracy is still statistically significant, $r = .31$. Although the observed improvement in fit due to inclusion of the a parameter is fairly modest, this analysis suggests that this parameter does capture meaningful differences among individuals in their memory for training trial information.

The other imperfect memory model, allowing encoding and retrieval errors in which cue presence and absence are reversed and incorrectly associated with the

Table 2c
Estimated parameter values and index of fit of 4-parameter ESAM assuming perfect memory compared with that of alternative models assuming imperfect memory

Model	Experiment	α	β	γ	δ	a/b	RMSE	t
ESAM	1	—	0.10	2.96	0.22	—	4.82	
	2	—	0.66	1.14	0.27	—	6.19	
	3	—	0.32	1.62	0.27	—	6.85	
	4	0.25	0.28	2.52	0.22	—	6.88	
Imperfect	1	—	0.10	2.96	0.22	1.00	4.82	0.00
memory	2	—	0.31	2.12	0.37	0.70	6.00	−1.67*
(Hypothesis)	3	—	0.17	2.69	0.37	0.72	6.76	−3.31***
	4	0.23	0.26	2.67	0.24	0.98	6.81	−2.83***
Imperfect	1	—	0.10	2.96	0.22	1.00	4.82	−0.20
memory	2	—	0.33	2.03	0.36	0.78	6.00	−1.71*
(abs/pres)	3	—	0.17	2.69	0.37	0.80	6.76	−3.56***
	4	0.25	0.28	2.53	0.22	1.00	6.88	−2.56**

Note. α, baserate adjustment; β, enhanced discounting; γ, extremity; δ, absent cue weighting; a/b, memory accuracy; RMSE, root mean squared error of ESAM's fit to the set of 192 mean probability judgments (on the 0–100% judgment scale) in each experiment; t, paired t test comparing fit of alternative model to that of ESAM with equal number of free parameters, based on individual judgment data.

*$p < .10$.
**$p < .05$.
***$p < .01$.

Table 3
Median estimated ESAM parameter values and median index of model fit to individual judgment data from Experiments 1–4

Experiment	α	β	γ	δ	RMSE
1	—	0.08	2.88	0.16	24.0
2	—	0.77	1.16	0.29	28.7
3	—	0.31	1.58	0.25	30.4
4	0.24	0.29	2.55	0.23	24.5

Note. RMSE, median root mean squared error of ESAM's fit to an individual's set of 192 probability judgments (on the 0–100% judgment scale).

Table 4
Comparison of ESAM and unnormalized diagnosticity model fit to probability judgment data from Koehler (2000) Experiment 3, along with estimated parameter values

Model	α	β	γ	δ	RMSE
ESAM	—	0.68	1.38	0.24	8.75
Unnormalized diagnosticity	—	0.41	1.39	0.21	9.29

Note. RMSE, root mean squared error of ESAM's fit to the set of 96 mean probability judgments (on the 0–100% judgment scale) of Koehler (2000) Experiment 3.

relevant hypothesis, yielded nearly identical results (see Table 2c). One might suspect that our exclusion from the data of participants with poor performance in the training phase of the experiment might have played a role in this result; that is, perhaps the imperfect memory models would exhibit a more pronounced advantage over the original ESAM model had we included the judgment data from these participants in the analysis. In investigating this possibility, however, we found that the small advantage of the imperfect memory models over the original version of ESAM that assumes perfect memory did not change much in magnitude even when the poor-performing participant data were included in the model-fitting comparison.

Taken together, the results from fitting the imperfect memory models suggest that, at least in the experimental context we investigated, ESAM's perfect memory assumption may often be a useful simplification. The imperfect memory versions of ESAM may be more appropriate under less ideal learning conditions as, for example, when training takes place under conditions of divided attention or when the number of cues or hypotheses increases to the point where encoding and retrieval errors are more prevalent.

7. Generalization of the Bayesian model

As noted earlier, because of several key differences in form between ESAM and the Bayesian model, there are no parameter values for which ESAM will reproduce the corresponding Bayesian values exactly. This makes it somewhat difficult to evaluate parameter estimates (in particular, the values of α and γ) in ESAM from the

normative perspective offered by the Bayesian model. In this section, as a means of addressing this potential drawback, we introduce a generalization of the Bayesian model (which we will refer to as Bayesian ESAM) using the same basic principles (and corresponding parameters) used to develop the original version of ESAM. Earlier it was suggested that any model developed from the principles of evidential support assessment outlined in Section 2 would produce a reasonably close fit to the data. The Bayesian generalization offers a test of this claim, and yields a variant of ESAM that might be particularly useful when judgment data are fit with an emphasis on their accuracy or correspondence to the output of a Bayesian analysis.

We start with the Bayesian model (14) assuming—as does ESAM—conditional independence of cues. In the Bayesian model, the value corresponding to the support for a hypothesis is given by

$$s_{\mathbf{C}}(H) = f(H)\prod_{i=1}^{N_C}\frac{f(C_i,H)}{f(H)} \text{ over cues } C_i \text{ in cue pattern } \mathbf{C}. \tag{18}$$

If the support for each hypothesis is calculated as in (18), the (Bayesian) probability of a hypothesis is given by its support normalized relative to that of its alternatives.

In the Bayesian version of ESAM, the support for a hypothesis is given by

$$s_{\mathbf{C}}(H) = \left[f(H)^{\alpha}\prod_{i=1}^{N_C}\frac{\delta_i f(C_i,H)}{f(H)}\right]^{\gamma} \text{ over cues } C_i \text{ in cue pattern } \mathbf{C}, \tag{19}$$

where $\delta_i = 1$ for present cue values and $\delta_i = \delta$ for absent cue values. The free parameter δ is expected to be less than 1 if absent cue values are underweighted. The free parameters α and γ, as in the original version of ESAM, reflect the relative weight placed on the baserate of the hypothesis under evaluation and the extremity of the resulting support values, respectively. If the baserate of the hypothesis is underweighted relative to the diagnostic value of the evidence (i.e., cue pattern), then $\alpha < 1$. If judgments based on the available evidence are more or less extreme than is justified according to the Bayesian analysis, then γ will tend to be greater or less than 1, respectively. The support calculation (19) is equivalent to that of the Bayesian model when $\alpha = \gamma = \delta = 1$.

The assumption of enhanced residual discounting is incorporated in Bayesian ESAM by discounting the contribution of each cue value according to its probability of being consulted, where the discounting weight $q_{\overline{H}}$ is calculated as in the original version of ESAM (12) using the free parameter β. The support for the residual hypothesis $\overline{H}$ is then given by

$$s_{\mathbf{C}}(\overline{H}) = \sum_{H_j \text{ in } \overline{H}}\left[f(H_j)^{\alpha}q_{\overline{H}}\prod_{i=1}^{N_C}\frac{\delta_i f(C_i,H_j)}{f(H_j)}\right]^{\gamma}. \tag{20}$$

When $\beta = 0$ (i.e., $q_{\overline{H}} = 1$), the model produces additive judgments as in the Bayesian approach. Positive values of β indicate that support for alternative hypotheses is discounted to a greater extent, producing increasingly subadditive judgments, as support for the focal hypothesis increases. The resulting model, then, is a generalization of the

Bayesian approach, based on the principles used to develop ESAM, that reduces to the Bayesian model when $\alpha = 1$, $\beta = 0$, $\gamma = 1$, and $\delta = 1$.

Because of the common underlying principles on which they are based, we expected that Bayesian ESAM would fit the data from our experiments about as well as the original version of ESAM. Bayesian ESAM's performance was indeed extremely close to that of the original version of ESAM. Table 2a indicates that, on average, neither model fits the data better than the other, $F(1, 133) = 0.87$, n.s. The reason for their similar performance, of course, is that the two models produce highly correlated predictions, as would be expected given their shared roots. The 192 predicted judgments made by the two models when fit to the mean judgments from each study, for example, were highly correlated with one another ($r = .98$ across all four experiments). Furthermore, the best-fitting parameter values for an individual participant were highly correlated across the original and Bayesian versions of ESAM, with $r = .73$ for α, $r = .73$ for β, $r = .97$ for γ, and $r = .37$ for δ.

Not surprisingly, predictions from the Bayesian ESAM model correlate more highly with the corresponding Bayesian values than do those of ESAM: with $r = .97$, .92, .93, and .96 in Experiments 1–4, respectively, as opposed to $r = .94$, .88, .88, and .93, respectively, for the original version of ESAM. As was the case for the original version of ESAM, Bayesian ESAM is also able to account for substantial variance in the observed judgments that is not accounted for by the standard Bayesian model. The partial correlation between Bayesian ESAM's predictions and the observed mean judgments, controlling for the corresponding Bayesian values, is .62, .50, .54, and .58 in Experiments 1–4, respectively.

Bayesian ESAM is also able to reproduce the tendency for T to increase with the number of present cues in the cue pattern. This may seem somewhat surprising given that, as pointed out in our earlier discussion of the Bayesian model, the product form of the integration method in (18) yields decreasing support as the number of cues consulted increases. Recall that ESAM reproduces the pattern of increasing T because support for the focal hypothesis increases more quickly than support for the alternative hypotheses as the number of present cue values in the cue pattern increases. Bayesian ESAM, by contrast, reproduces the pattern of increasing T because support for the focal hypothesis decreases less quickly than does support for the alternative hypotheses as the number of number of present cue values in the cue pattern increases.

The free parameter estimates that result from fitting Bayesian ESAM to the judgments data from the four experiments are largely consistent with the theoretical account upon which the original version of ESAM was based. The tendency to underweight cue absence is reflected in estimates of δ being consistently less than 1. Enhanced residual discounting is reflected in values of β that are consistently greater than 0. The construction of Bayesian ESAM as a generalization of the Bayesian approach provides additional interpretability of the α and γ parameter estimates. The observation of $\alpha < 1$ indicates that the baserate of the hypothesis is underweighted in the support assessment process relative to the requirements of the standard Bayesian model. The observation of $\gamma < 1$ across all four experiments indicates

that the observed judgments were less extreme than would be those of a true Bayesian, reflecting a tendency toward conservatism (Edwards, 1968).

Given that the estimated parameter values from the original and Bayesian version of ESAM are highly correlated, it is not surprising that the relationships found between the best fitting parameters in ESAM and certain other measures of individual differences in judgments also hold for the Bayesian version of ESAM. For instance, the estimated value of β for an individual in Bayesian ESAM is highly correlated with his or her mean value of T; after removing the individuals who were outliers on this parameter, partialling out the effects of experiment and controlling for the value of the γ and δ parameters, $r = .64$, $p < .001$. And once again, the estimated value of γ correlated with a participant's choice accuracy during the training phase of the experiment; after partialling out the effects of experiment and controlling for the value of the β and δ parameters, $r = .48$, $p < .001$.

8. Conclusions and remaining questions

Despite its simplicity, ESAM was found to provide a reasonably good fit to the observed judgments in four experiments. Alternative models differing from ESAM by one or more key assumptions tended to fit the data less well. Specifically, models that consider entire cue patterns rather than single cues, models that fail to place greater weight on cue presence than on cue absence, models that use different methods for calculating the diagnostic value of a cue, models that do not adjust for the baserate of a hypothesis, and models that assume no relationship between the support for the focal hypothesis and discounting of support for its alternatives, all fit the data less well than ESAM. 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 experimental data. The specific details of our implementation of ESAM are less important than the general conclusion that evidential support assessment in probability judgment can be characterized as a process in which the diagnostic implications of a body of evidence are evaluated on a cue-by-cue basis in light of previous experience with those cues. The comparable performance of a variant of ESAM based on a generalization of the Bayesian model is consistent with this view.

For the experiments we reported, ESAM tended to produce reasonably accurate judgments 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 (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' rule, integrates considerations of the diagnosticity of the evidence and of the baserate or prior probability of the hypothesis. The observation that

ESAM and the Bayesian approach produce highly correlated output 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).

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 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. This assumption is shared by other recent models of classification learning and probability judgment, including an adaptive network model developed by Gluck and Bower (1988). An alternative approach is offered by exemplar-based models, such as Medin and Schaffer's (1978) context model and Kruschke's (1992) ALCOVE, which presume storage and evaluation of entire cue patterns (i.e., exemplars).

Exemplar-based models 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 our experiments, counts must be maintained of the frequency with which each of the 64 possible cue patterns co-occur with each hypothesis in an exemplar representation. The prototype representation requires only 12 counts (i.e., an absence and a presence count for each cue) per hypothesis, rather than 64, 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. One benefit of retaining additional information in the exemplar representation, as discussed previously, is that it allows use of configural information in the case of conditionally dependent cues. Another benefit is flexibility: The counts maintained in ESAM's prototype representation require designation of an outcome variable (i.e., hypothesis) around which the frequency counts are organized; a separate set of counts would be required if the designation of one variable as the outcome and another as a cue were to be reversed. The exemplar representation, by contrast, allows reorganization around any variable that might be designated as the outcome of interest because it maintains information on the co-occurrence of all variables under evaluation.

The formulation of ESAM as taking as input information in a prototype representation does not, of course, preclude the possibility that an exemplar-based representation is also maintained in memory. An exemplar representation could be used as the starting point, for instance, in calculating the individual cue-hypothesis co-occurrence frequencies that ESAM takes as input. The main point of our results is merely that ESAM can produce what turn out to be reasonably accurate judgments

on the basis of a relatively small number of stored frequency counts using a prototype representation.

That said, it is also possible, at least in complex uncertain environments characterized by a large number of diagnostic cues, that memory for exemplar frequencies is not stored directly but instead is reconstructed from individual cue-outcome co-occurrence frequencies. As might be expected if this were the case, Koehler (2000) observed a pattern of systematic subadditivity and enhanced residual discounting in judgments of absolute frequency (of exemplars) in a task very similar to that of the present experiments. In that task, participants given a particular symptom pattern first estimated how many patients with that exact pattern they had seen during training, and second estimated of those how many had a designated flu strain. The resulting proportions were slightly less subadditive than a corresponding set of probability judgments from another experiment, but both exhibited consistent subadditivity that increased with the number of present symptoms in the symptom pattern. The model developed in this paper can be used to capture such influences on judgments of the frequency with which various exemplars have been observed, on the assumption that these judgments are in turn based on accurate frequency counts of the co-occurrence of individual cues with the hypotheses of interest.

In light of the accuracy it achieves despite numerous simplifying assumptions, ESAM might be viewed as belonging to a larger family of judgmental heuristics, introduced by Tversky and Kahneman (1974), that often produce reasonably accurate judgments but that also exhibit systematic biases under certain conditions. In contrast to the work of Tversky and Kahneman, however, which emphasizes judgments derived from "natural assessments" (Tversky & Kahneman, 1983) of similarity or availability, the process by which ESAM translates stored frequency counts into assessments of evidential support can perhaps be more precisely described as a cognitive algorithm. Recently, a number of researchers, most notably Gigerenzer and his colleagues (Gigerenzer, Todd, & The ABC Research Group, 1999), have explored similar algorithms that produce acceptably accurate judgments, which they describe as "fast and frugal," reflecting their goal that these algorithms should be as computationally simple and informationally undemanding as possible. Gigerenzer and colleagues have focused much of their attention on one particular algorithm called "Take the Best" (TTB; Gigerenzer & Goldstein, 1996), the origin of which lies in their earlier research on probabilistic mental models of judgmental confidence (Gigerenzer, Hoffrage, & Kleinbölting, 1991). Because of its prominence, it may be instructive to compare TTB to ESAM, even though the models were developed to carry out somewhat different tasks.

The task of TTB is to determine which member of a pair of judgment targets (e.g., cities) has the higher value on a given dimension (e.g., population) and also, in the original probabilistic mental models formulation, to assign a probability that the resulting judgment is correct. The most obvious similarity between TTB and ESAM is that both rely on (perfectly accurate) stored counts summarizing the frequency with which various cues have been observed to co-occur with some outcome variable of interest to the judge. In TTB, the frequency counts are used to determine and rank

order the cues by their predictive validity. Then, to choose the pair member judged to have the higher value on the dimension in question, the cues are consulted one at a time, in order of decreasing predictive validity, until one is found for which the two pair members have different values. The cue thus selected is used to make the choice between the pair members, and the judged probability of the choice being correct is given by the validity of that cue.

In addition to their being developed for different judgmental tasks, the key differences between TTB and ESAM are as follows:

1. TTB consults only a subset of the available cues in arriving at a judgment; ESAM consults all the available cues in assessing the support for the focal hypothesis, and consults a subset of the cues in assessing the support for alternatives included in the residual. TTB consults a smaller number of cues when the members of the pair of judgment targets are highly discriminable (i.e., share fewer attributes or cue values) than when they are less discriminable; ESAM consults a smaller number of cues in assessing the support for alternative hypotheses included in the residual as the support for the focal hypothesis increases.

2. TTB renders a decision and a probability estimate on the basis of a single cue and therefore does not have to integrate the implications of multiple cues; ESAM integrates the independently assessed implications of each in the set of cues consulted (i.e., assuming conditional independence of cue values). In other words, TTB is a noncompensatory model while ESAM is a (partially) compensatory model.

3. TTB requires substantial "precomputation" of the stored frequency counts to determine and rank order the cues by their predictive validity, which is defined as the proportion of cases, out of all possible pairs of judgment targets, that the cue correctly indicates the pair member with the higher value on the dimension in question; ESAM can be viewed as carrying out all its computations on the stored frequency counts at the time of judgment. That said, as formulated, ESAM has no mechanism for narrowing a set of potentially useful cues to a manageable number; its diagnostic value calculation would seem a natural starting point for such a process, but such a possibility will have to be explored in future work.

4. TTB produces judgments that are unbiased and additive; ESAM produces judgments that are generally subadditive, with the degree of subadditivity being predictable from characteristics of the cue pattern on which the judgments are based. TTB also requires supplementary assumptions to model judgments concerning more than two competing hypotheses (e.g., Berretty, Todd, & Martignon, 1999).

Despite their differences, both TTB and ESAM exemplify growing interest in the development of simple models that produce accurate judgments. Whether such models describe how people actually make probability judgments is another question. The preliminary evidence reported here indicates that ESAM is promising in this regard; to date, relatively little effort has been put into comparing the predictions of TTB to observed probability judgments, with at best equivocal results from recent work on this question (Bröder, 2000; Rieskamp & Hoffrage, 1999; Slegers, Brake, & Doherty, 2000), although in this regard it should be noted that TTB has no free parameters and as such it is somewhat complicated to directly compare its "fit" to that of models such as ESAM. Clearly, further research will be necessary to more

fully evaluate the adequacy of such models as descriptive accounts of human judgment.

While ESAM as it is currently implemented appears to hold some promise, it will require more extensive testing and possible revisions in light of that testing. First, the model could usefully be extended to cases of missing information, that is, circumstances in which only a subset of the relevant cue values are known to the judge. Second, it will be necessary to investigate whether or not ESAM's performance generalizes to cases in which conditional independence of cues does not hold. It seems implausible that people will entirely ignore obvious conditional dependencies among cues, as implied by the prototype representation employed by the current version of ESAM. Instead, people may employ configural judgment strategies in which they focus on the implications of clusters of conditionally dependent cue values (e.g., Edgell, 1978, 1980). In such cases, a better understanding of how people segregate cues into roughly independent clusters may allow application of the model to a wider variety of judgmental settings.

References

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

Berretty, P. M., Todd, P. M., & Martignon, L. (1999). Categorization by elimination: Using few cues to choose. In G. Gigerenzer & P. M. Todd et al. (Eds.), *Simple heuristics that make us smart* (pp. 235–254). Oxford: Oxford University Press.

Birnbaum, M. H., & Mellers, B. A. (1983). Bayesian inference: Combining base rates with opinions of sources who vary in credibility. *Journal of Personality and Social Psychology, 45*, 792–804.

Brenner, L. A., & Koehler, D. J. (1999). Subjective probability of disjunctive hypotheses: Local-weight models for decomposition of evidential support. *Cognitive Psychology, 38*, 16–47.

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

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: Erlbaum.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120*, 278–287.

Brunswik, E. (1956). *Perception and the representative design of experiments*. Berkeley, CA: University of California Press.

Castellan, N. J., Jr. (1977). Decision making with multiple probabilistic cues. In N. J. Jr., Castellan, D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. II, pp. 117–147). New Jersey: Erlbaum.

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

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

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

Edwards, W. E. (1968). Conservatism in human information processing. In B. Kleinmutz (Ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83*, 37–64.

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology, 18*, 500–549.

Estes, W. K., Campbell, J. A., Hatsopoulos, N., & Hurwitz, J. B. (1989). Base-rate effects in category learning: A comparison of parallel network and memory storage-retrieval models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*, 556–571.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1978). Fault trees: Sensitivity of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance, 4*, 330–344.

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

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

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117*, 227–247.

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist, 39*, 1372–1388.

Hutchinson, J. W., & Alba, J. W. (1991). Ignoring irrelevant information: Situational determinants of consumer learning. *Journal of Consumer Research, 18*, 325–345.

Kao, S.-F., & Wasserman, E. A. (1993). Assessment of an information integration account of contingency judgment with examination of subjective cell importance and method of information presentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1363–1386.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 317–330.

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.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Kruschke, J. K., & Johansen, M. K. (1999). A model of probabilistic category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1083–1119.

Martignon, L., & Laskey, K. B. (1999). Bayesian benchmarks for fast and frugal heuristics. In G. Gigerenzer & P. M. Todd et al. (Eds.), *Simple heuristics that make us smart* (pp. 169–188). Oxford: Oxford University Press.

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

Meyer, R. J. (1987). The learning of multiattribute judgment policies. *Journal of Consumer Research, 14*, 155–173.

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

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

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.

Peterson, C. R., Hammond, K. R., & Summers, D. A. (1965). Optimal responding in multiple-cue probability learning. *Journal of Experimental Psychology, 70*, 270–276.

Rieskamp, & Hoffrage, (1999). When do people use simple heuristics, and how can we tell. In G. Gigerenzer & P. M. Todd et al. (Eds.), *Simple heuristics that make us smart* (pp. 141–167). Oxford: Oxford University Press.

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

Schustack, M. W., & Sternberg, R. J. (1981). Evaluation of evidence in causal inference. *Journal of Experimental Psychology: General, 110*, 101–120.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8*, 208–224.

Slegers, D. W., Brake, G. L., & Doherty, M. E. (2000). Probabilistic mental models with continuous predictors. *Organizational Behavior and Human Decision Processes, 81*, 98–114.

Smedslund, J. (1963). The concept of correlation in adults. *Scandinavian Journal of Psychology, 4*, 165–173.

Trope, Y., & Mackie, D. M. (1987). Sensitivity to alternatives in social hypothesis-testing. *Journal of Experimental Social Psychology, 23*, 445–459.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Tversky, A., & Kahneman, D. (1983). Extensional vs. intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 91*, 293–315.

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

Van Osselaer, S. M. J., & Alba, J. W. (2000). Consumer learning and brand equity. *Journal of Consumer Research, 27*, 1–16.