Journal of Experimental Psychology:
Learning, Memory, and Cognition
2000, Vol. 26, No. 1, 28–52

0278-7393/00/$5.00 DOI: 10.1037//0278-7393.26.1.28

Probability Judgment in Three-Category Classification Learning

Derek J. Koehler
University of Waterloo

People give subadditive probability judgments—in violation of probability theory—when asked to assess each in a set of 3 or more mutually exclusive hypotheses, as indicated by their sum exceeding 1. Three potential evidential influences on subadditivity—cue conflict, cue frequency, and cue redundancy—are distinguished and tested in 5 experiments using a classification-learning task. Results indicate that (a) judgments of probability and of frequency are systematically subadditive even when the judgments are based on cues learned within the experimental context, (b) cue conflict has a reliable influence on the degree of subadditivity, and (c) judgments in this context are well described by a linear-discounting model within the framework of support theory.

New York Yankees catcher Yogi Berra was once asked whether he would like his pizza cut into four or eight slices, to which he is supposed to have replied that he would prefer it be cut into four slices, as he was not hungry enough to eat eight. Piaget and Inhelder (1941, cited in Flavell, 1963) observed that young children do in fact fail to understand that mass or quantity is conserved when an object (e.g., a piece of clay or a glass of water) is partitioned into components (e.g., several pieces of clay or multiple glasses of water). Although adults are not prone to make such fundamental errors in reasoning about physical quantities, on more complicated tasks they appear to retain an inclination to base judgments on the number of components into which the object or event is partitioned, even when the number of components is irrelevant with respect to the value being assessed (e.g., Fiedler & Armbruster, 1994; Pelham, Sumarta, & Myaskovsky, 1994; van der Pligt, Eiser, & Spears, 1987). This article concerns effects of event decomposition in judgments of likelihood. Subjective assessments of probability, contrary to the dictates of probability theory, are often highly dependent on the manner in which the event under assessment is partitioned into components.

Specifically, a number of studies have demonstrated that an event is assigned a greater probability when its components are explicitly listed and individually assessed than when evaluated as a whole (Dube-Rioux & Russo, 1988; Fischhoff, Slovic, & Lichtenstein, 1978; Fox, Rogers, & Tversky, 1996; Fox & Tversky, 1998; Mehle, Gettys, Manning, Baca, & Fisher, 1981; Peterson & Pitz, 1988; Teigen, 1974a, 1974b). For example, the judged probability of death due to homicide increases when this possibility is "unpacked" into homicide by an acquaintance or homicide by a stranger (Rottenstreich & Tversky, 1997). Likewise, physicians' estimates of the probability of alternatives to a focal diagnosis increase when a number of specific alternative diagnoses are explicitly mentioned (Redelmeier, Koehler, Liberman, & Tversky, 1995).

Recently, a descriptive theory of probability judgment called *support theory* (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) has been developed that can account for such findings. Support theory consists of two basic assumptions. The first is that judged probability reflects the relative support for the focal and alternative hypotheses:

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

That is, the judged probability of A rather than B is simply the evidential support available for A, $s(A)$, normalized relative to that available for its complement B. If, for example, A represents the possibility that a Democrat will win the next presidential election, and B represents the possibility that a Republican will win, the judged probability of a Democrat rather than a Republican winning $P(A, B)$ is assumed to be represented as the proportion of evidential support $s(A)$ for the Democrat relative to the evidential support $s(B)$ for the Republican. Support theory is nonextensional, allowing judged probability to depend not only on the event in question but also on how it is described. Hence, A

Experiment 1 was conducted during a postdoctoral visit at the Medical Research Council Applied Psychology Unit in Cambridge, England, and was funded by the National Science Foundation's Program for Long- and Medium-Term Research at Foreign Centers of Excellence. Experiment 2 was conducted at University College London and was supported by Grant R000221383 from the Economic and Social Research Council of the United Kingdom. I thank Alan Baddeley and Nigel Harvey for acting as gracious hosts during my visits to these two institutions, respectively. Experiments 3–5 were conducted at the University of Waterloo and supported by Grant OGP 0183792 from the Natural Sciences and Engineering Research Council of Canada.

Thoughtful suggestions and comments on this work were generously provided by Lyle Brenner, Steve Edgell, Dale Griffin, Stephen Lewandowsky, Barbara Malt, Robert Nosofsky, David Shanks, and the late Amos Tversky. I am grateful to Emily Marks and Jodi Cryderman for their assistance in conducting several of the experiments reported in this article.

Correspondence concerning this article should be addressed to Derek J. Koehler, Department of Psychology, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada. Electronic mail may be sent to dkoehler@watarts.uwaterloo.ca.

and B refer to descriptions of events, called *hypotheses*, rather than to the events (in the set-theoretic sense) themselves, as in standard probability theory.

Support theory distinguishes between 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., Hurricane Bonnie will come ashore along the eastern U.S. coastline) that refers to the same event as an explicit disjunction of exclusive hypotheses A_1 and A_2 (e.g., Hurricane Bonnie will come ashore along the northeastern U.S. coastline or Hurricane Bonnie will come ashore along the southeastern U.S. coastline, denoted $A_1 \vee A_2$), then

$$s(A) \leq s(A_1 \vee A_2) \leq s(A_1) + s(A_2). \quad (2)$$

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 total 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. As with Berra's pizza, decomposition increases perceived extent or likelihood. The observed effects of unpacking reported in numerous studies (for a review, see Tversky & Koehler, 1994) are inconsistent both with the standard Bayesian model of subjective probability and with nonstandard models such as Shafer's (1976) theory of belief functions.

Support theory implies that, whenever an elementary hypothesis is evaluated relative to all of its alternatives taken as a group (referred to as a "catchall" or *residual* category), 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 student is known to major in one (and only one) of three possible social sciences: economics (A), psychology (B), or sociology (C). According to support theory, when a person is asked to judge the probability of Hypothesis A (i.e., that the student majors in economics rather than psychology or sociology), the resulting "elementary" probability judgment is determined by the evidential support for Hypothesis A normalized relative to that for its complement (not-A, represented $\overline{A}$). In this case, its complement is an implicit disjunction of Hypotheses B and C. Support theory assumes that packing these alternatives together in the implicit disjunction (i.e., the residual) generally produces a loss in their support, thereby increasing A's judged probability.

If separate elementary judgments are obtained of the probability of hypotheses A, B, and C, the total probability T assigned to the three elementary hypotheses is predicted to exceed one, in violation of probability theory. This result is predicted only when there are three or more elementary hypotheses under evaluation, as it is only under these circumstances that the alternatives to the focal hypothesis can be packed together to create an implicit residual hypothesis. (In the case of a complementary pair of hypotheses, judgments are predicted to sum to one; see Equation 1. For some exceptions, see Brenner & Rottenstreich, 1999; Macchi, Osherson, & Krantz, 1999; McKenzie, 1998, 1999.) The degree of subadditivity 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. (Unfortunately, the term *subadditivity* can be somewhat confusing in this context, given that it is indicated by $T > 1$. This terminology is supposed to reflect the fact that an event as a whole receives less probability than the sum of that assigned to its components.)

The observed degree of subadditivity depends on a number of factors (see Tversky & Koehler, 1994), including the compatibility of the evidence with the set of hypotheses under consideration. For example, in one experiment (Koehler, Brenner, & Tversky, 1997, Experiment 1) participants judged the probability that a college student had a specified social science major on the basis of a course that student had taken. The courses provided as evidence varied in how compatible they were with social science majors in general, with two of them being quite typical (e.g., Western Civilization) and two being fairly atypical (e.g., French Literature). The degree of subadditivity of the judgments, as measured by the total probability T assigned to four exclusive and exhaustive social science majors, was significantly greater for the typical courses than for the atypical courses, a result referred to as the *enhancement effect* (Brenner & Koehler, 1999; Koehler et al., 1997; Tversky & Koehler, 1994). Apparently, when evidence is introduced that is generally supportive of each of the hypotheses under evaluation, the focal hypothesis of the elementary judgment receives a disproportionate share of the perceived support conveyed by the evidence.

While the notion of compatibility between evidence and hypotheses serves to summarize a number of evidential manipulations observed to influence subadditivity, the specific characteristics of the evidence that contribute to subadditive judgments have yet to be explicated. To identify more precisely the evidential characteristics influencing subadditivity, it is necessary to have direct experimental control over the relationship between the evidence and the hypotheses. In the experiments reported in this article, this was accomplished through the use of a simulated medical diagnosis task, a well-established experimental paradigm that has been used in much of the recent work on classification learning (or, more precisely, multiple-cue probability learning; for a review of much of the early work in this area, see Castellan, 1977).

In these studies (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Nosofsky, Kruschke, & McKinley, 1992; Shanks, 1991), participants are presented with a set of symptoms ("cues" which serve as evidence) reported by a "patient" and are asked to diagnose which of a set of possible diseases (typically two) the patient might have. Participants are presented with a large number of patients; after making each diagnosis, participants receive

feedback telling them which disease the patient actually had. During or after the learning phase, test trials may be given (typically without feedback), in which participants are presented with symptom patterns and asked to estimate the probability that the patient has a designated disease.

The present experiments use this kind of simulated medical diagnosis task to investigate three evidential factors that could potentially influence the degree to which probability judgments are subadditive. While support theory offers a framework for interpreting the impact of these factors, it does not directly predict that any of these factors will necessarily influence the degree of subadditivity observed in probability judgments. These manipulations, then, are intended to further our understanding of how evidential support is assessed, not to provide a direct test of support theory itself.

The first factor, referred to as *cue conflict,* reflects the extent to which the evidence is "mixed" in its implications (cf. Peterson & Pitz, 1988). Certain symptom patterns may include some symptoms that are supportive of one diagnosis and other symptoms that are supportive of a different diagnosis, inducing a high state of conflict; other symptom patterns, by contrast, may provide support for only a single diagnosis, thereby inducing less conflict. One interpretation of previous studies of the enhancement effect (e.g., Koehler et al., 1997) is that the experimental manipulations used in these studies operated by varying cue conflict. Under this interpretation, a greater degree of subadditivity should be associated with increased cue conflict.

The second factor, referred to as *cue frequency,* concerns the general prevalence of a cue's presence in the learning environment. Certain cues may just be more commonly encountered than other cues, independent of their diagnostic value. The present experiments investigate the possibility that high-frequency cues induce greater subadditivity than low-frequency cues. This might occur, for example, if people attend to the degree of co-occurrence between a cue and a particular diagnosis, without due regard to its co-occurrence with alternative diagnoses (cf. Jenkins & Ward, 1965).

The symptoms acting as cues in the reported experiments are all binary in nature, and are denoted simply as either present or absent (e.g., *cough* vs. *no cough*). As will become more apparent when the experiments are described, an underlying assumption in evaluating the impact of cue conflict and cue frequency is that participants focus on and draw inferences primarily on the basis of a symptom's presence rather than its absence. Thus, cue conflict is assessed in terms of the number of diagnoses implicated by symptoms reported as present, and cue frequency is computed in terms of the prevalence of symptoms present in the evidence upon which the judgment is based. This approach implicitly assumes that participants conceptualize the cues in terms of what a symptom's presence—rather than its absence—signifies regarding the patient's condition. Obviously, this does not imply that participants fail to distinguish between present and absent symptoms (in which case they would exhibit no learning). Instead, the assumption is that participants represent information they acquire about the category structure (i.e., symptom–diagnosis interrelationships) primarily in terms of what a symptom's presence indicates about the patient's likely diagnosis.

There is some empirical evidence to support this assumption. For example, Estes et al. (1989) and Shanks (1990) report that participants in their classification-learning studies give judgments that generally fail to distinguish between absence of information about a symptom and information that a symptom is definitely absent. Such a pattern of results would be expected if participants base their judgments primarily on present symptoms. Research on judgments of covariation and causation (e.g., Kao & Wasserman, 1993; Schustack & Sternberg, 1981; Shaklee & Mims, 1982; Smedslund, 1963) also indicates 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. Generally speaking, people appear to be more sensitive to the diagnostic value of a cue's presence than to the diagnostic value of its absence (Newman, Wolff, & Hearst, 1980; Norton, Muldrew, & Strub, 1971). One reason, of course, why greater weight may be placed on a cue's presence than on its absence is that the physical presence of a cue draws attention while its absence does not. Even when a cue's absence is explicitly denoted (e.g., by providing a 2×2 frequency table in contingency judgment studies, or by explicitly denoting a symptom's absence in a patient description in a classification learning task), however, greater weighting of cue presence than cue absence can still be observed (Estes et al., 1989; Kao & Wasserman, 1993).

This issue will be considered further in the general discussion. For now, the main point is methodological: Effects of evidential factors such as cue conflict and cue frequency are assessed in this article with respect to symptoms that are said to be present rather than absent. Such a focus draws attention to a third potential factor, referred to as *cue redundancy.* Holding constant the diagnostic implications of a pattern of symptoms (and thus the degree of cue conflict), it is possible that the mere presence of additional (nondiagnostic) cues will produce increased subadditivity. This factor is investigated in Experiments 3–5.

In all five experiments, participants were presented with computer-simulated "patients," each of whom was said to suffer from one (and only one) of three possible flu strains. Each experiment consisted of a learning phase followed by a judgment phase. In the learning phase, participants diagnosed the flu strain a patient was suffering from on the basis of a set of symptoms said to characterize that patient. Feedback regarding the correct diagnosis was presented, so that over the course of the learning phase, participants acquired a sense of which symptoms were associated with which flu strains. In the judgment phase, which is the main focus of this investigation, participants assessed the probability that a patient, characterized by a particular symptom pattern, was suffering from a designated flu strain. Because the three flu strains were known to be mutually exclusive and collectively exhaustive, probability theory requires that the total probability T assigned to the three flu strains given a particular symptom pattern add to 1 (or 100%). Support

theory, in contrast, implies subadditive judgments such that T will generally exceed 1.

Experiment 1 confirms this general prediction and allows assessment of the impact of cue conflict and cue frequency on subadditivity as measured by the value of T. Experiment 2 investigates whether providing feedback after every probability judgment reduces or eliminates the tendency toward subadditive judgments. Experiment 3, which uses a different category structure relating symptoms to flu strains, establishes the generalizability of the results and allows examination of the effect of cue redundancy in addition to that of cue conflict and cue frequency. Experiment 4 demonstrates that the same general pattern of results holds for judgments of frequency as well as of probability. Finally, Experiment 5 was conducted to collect some supplementary data that would allow the fitting of a linear-discounting model of support (Koehler et al., 1997) to the results of Experiments 3 and 4. Taken together, these experiments provide some insight into the factors influencing the perceived evidential support for a hypothesis, using support theory as a guiding theoretical framework.

Experiment 1

The first experiment examines the influences of cue conflict and cue frequency. For each patient in the simulated medical diagnosis task, participants were provided with information regarding four symptoms which varied in their overall frequency. During the learning phase, participants chose one of the three flu strains as their diagnosis on each trial, and then received feedback regarding the correct diagnosis. During the judgment phase, participants were presented with symptom patterns and asked to judge the probability of a designated flu strain given that pattern. Some of the symptom patterns included multiple symptoms associated with different flu strains, inducing a state of high cue conflict, while other symptom patterns included symptoms implicating only a single flu strain, inducing a state of low cue conflict. The total probability T assigned to the three flu strains given a particular symptom pattern is predicted to systematically exceed 1, and to increase with cue conflict and cue frequency.

Method

Participants. Participants were 16 members of the participant panel at the Medical Research Council Applied Psychology Unit, who were paid for their participation. Data from three additional participants were replaced; one participant failed to complete the judgment task appropriately, and the other two failed to achieve above-chance accuracy in the learning phase of the experiment.

Stimuli and apparatus. The stimuli were "medical charts" consisting of four symptoms: chills, cough, headache, and sore throat. Each symptom was denoted either as being present (in which case the symptom was written in uppercase letters, e.g., COUGH) or absent (in which case the symptom was written in lowercase letters, e.g., no cough) on the medical chart. The symptoms appeared in a vertically arranged list presented on a Macintosh computer, which was also used to record participants' judgments and provide feedback about the correct diagnosis for each patient.

Design. As in Estes et al. (1989) and Nosofsky et al. (1992), all participants were presented with an identical training sequence, consisting of 240 "patients" or trials. Each patient was to be classified as having one of three types of influenza (flu) strains, simply labeled Flu Strain 1, 2, or 3. The training sequence was constructed by first randomly choosing one of the three flu strains (with equal probabilities) and then choosing the four symptoms—independently—with the appropriate conditional probabilities for that flu strain. For Flu Strain 1, the probabilities of the symptoms being present were 1.00, 0.275, 0.225, and 0.1625 for Symptoms A–D respectively. For Flu Strain 2, the corresponding probabilities were 0.3375, 0.8375, 0.225, and 0.1625. For Flu Strain 3, the corresponding probabilities were 0.3375, 0.275, 0.6625, and 0.50. These probabilities yield the following properties. First, the four symptoms vary systematically in their overall frequency of occurrence, with $p(A) = 55.8\%$, $p(B) = 46.3\%$, $p(C) = 37.1\%$, and $p(D) = 27.5\%$. Second, the presence of each symptom, taken on its own, has the same diagnostic value. That is, given the symptom's presence, the flu strain it is associated with increases in probability to 60% (with some small variation due to rounding error for the finite series of training trials), and the other two flu strains decrease in probability to 20% each. Conditional probabilities given the absence of a symptom differ from one symptom to the next, with the associated flu strain's probability decreasing to 0%, 10%, 18%, and 23% and the remaining flu strains each increasing in probability to 50%, 45%, 41%, and 39%, given the absence of Symptoms A, B, C, and D, respectively.

The training sequence was constructed using these probabilities, with the additional constraints that (a) each of the three flu strains appeared on exactly one third of the training trials, (b) the targeted relative frequency of each symptom given each of the flu strains was achieved exactly over the training sequence, and (c) no symptom pattern appeared twice in succession anywhere in the sequence. The actual symptom (e.g., cough) assigned to the four abstract Symptoms A–D was counterbalanced over participants, as was the position of the four symptoms in the computer display. The training sequence was presented in a fixed order to all participants.

Following the training sequence, probability judgments were elicited in which the probability of a designated flu strain was estimated given a particular symptom pattern. Given four binary symptoms, there are 16 possible symptom patterns. When crossed with the three possible target flu strains, this produces a total of 48 possible "pattern judgments," the full set of which was obtained in a randomized order from each participant. The 48 symptom patterns vary both in their level of cue conflict and in the frequency of occurrence (during the training sequence) of the present symptoms they include. The level of cue conflict associated with a symptom pattern is represented by the number of flu strains implicated by the present symptoms it includes; cue frequency is represented by the average relative frequency (during the training sequence) of the present symptoms in the symptom pattern.

Procedure. Participants were told that they would be taking part in a simulated medical judgment task. In the first part of the experiment, they were told, they would be presented with a series of 240 patients, each of whom had come to the medical clinic with a body temperature greater than 100.5 degrees Fahrenheit and subsequently was found (via a blood test) to have one of three influenza strains (1, 2, or 3). They were instructed that their task was to consider four symptoms (chills, cough, headache, and sore throat) that could help them determine which of the three flu strains a patient was suffering from. For each patient, they would be told whether or not the patient had reported each of the four symptoms and then would be asked to guess which of the three flu strains that participant had. After entering their choice, they would be told whether they were correct or not and which flu strain the patient in

question actually had. In the beginning, they were told, they would only be guessing, but as they saw more patients they should begin to have some sense of which symptoms go with which flu strains. Participants were warned, however, that just as in real medical practice, these observable symptoms were not perfect predictors and that two patients with the exact same set of symptoms might not always have the same flu strain. Thus they were told not to expect to achieve perfect diagnostic accuracy even by the end of training sequence. They were informed that, after diagnosing the 240 patients, they would be asked to make some judgments regarding the relationship between the symptoms and the flu strains.

After the training sequence, participants were presented with symptom patterns (like those seen during training) and were asked to judge the percentage of patients with that pattern they would expect to have a designated flu strain. They were instructed to give numbers between 0% and 100%, where 100% indicated that they expected every patient with that symptom pattern to have the designated flu strain, and 0% indicated that they expected none of the patients with that pattern to have the designated flu strain. They were further told that they could think of their judgments as an indication of how certain they would be that a patient with the listed symptoms would have the designated flu strain. The instructions noted that the designation of a target flu strain would be made arbitrarily and hence should not be interpreted as a suggestion that it was particularly likely to characterize the patient in question.

Results and Discussion

Learning data. While the focus of the present research is on probability judgments made following learning, it is important to establish that participants did in fact learn something about the category structure during the training phase. Assessment of the learning data in terms of the percentage of correct choices over the 240 training trials provides a convenient measure of accuracy and allows identification of any participants who appeared to perform substantially worse than the typical participant. There are, of course, alternative measures of accuracy, but—given that the focus here is on probability judgments following learning—percent correct seems to offer a relatively straightforward, simple measure that is sufficient for current purposes.

Average accuracy across the 240 training trials was 55%, a proportion substantially greater than the 33% correct expected if participants were simply guessing on each trial. Individual participants varied considerably in their learning performance; the best performance was 64% correct, and the worst was 45% correct. All participants included in the analysis achieved above-chance accuracy during the training phase of the experiment.

To determine whether learning was at asymptote by the end of the 240 training trials, average percent correct (over participants) was computed for four consecutive 60-trial blocks. On the first block, 39% of participants' guesses were correct. For the next three blocks, the corresponding figures were 59%, 62%, and 58%, respectively. Apparently, participants' performance was no longer improving after the first 60 or so training trials, at least as measured by percentage of correct choices.

It is of interest to determine the theoretical maximum percent correct that could be achieved in this task, to get a sense of how well participants performed during learning. This was assessed by computing for each symptom pattern the most commonly associated flu strain in the training sequence and then evaluating the level of accuracy that could be achieved if that flu strain was offered as the guess on each presentation of the symptom pattern in question. Such a "maximizing strategy" yields an accuracy level of 72% across the 240 trials, suggesting that participants' performance on the last 180 trials was quite good but not at ceiling.

Pattern judgment data. Table 1 presents the mean probability judgments assigned to each flu strain for the 16 possible symptom patterns, with present symptoms denoted in uppercase and absent symptoms in lowercase. As expected, these probability judgments were clearly subadditive: The total probability T assigned to the three possible flu strains exceeded 100%, with an average total of 120%. Individually, only 2 of the 16 participants had average T values less than 100%; their averages were very nearly additive.

Participants' probability judgments were clearly influenced in the correct direction by the symptoms used as cues. The easiest way to see this is to examine the patterns that have only a single symptom present. For these patterns, participants gave high probabilities to the implicated flu strain (e.g., Flu Strain 1 was assigned a mean probability of 84% given pattern Abcd). To obtain a more complete measure of the accuracy of the mean judgments, one of two methods can be used to derive the "correct" probability values. First, the conditional probabilities used to construct the training sequence (i.e., the conditional probability of a symptom given a flu strain) can be combined using Bayes' rule to determine the expected value of a flu strain's probability given a particular symptom pattern. Because the training sequence was generated randomly and is of finite length, however, the actuarial value of a flu strain's relative

Table 1
Average Judged (Percent) Probability of Each Flu Strain for the 16 Possible Symptom Patterns, Along With Their Total T, *in Experiment 1*

Pattern	Flu 1	Flu 2	Flu 3	Total T
ABCD	19	22	64	105
ABCd	32	47	52	131
ABcD	43	42	47	132
ABcd	48	70	12	130
AbCD	34	25	83	142
AbCd	60	18	52	130
AbcD	52	20	47	119
Abcd	84	25	07	116
aBCD	13	45	73	131
aBCd	16	57	62	135
aBcD	14	60	46	120
aBcd	12	90	07	109
abCD	18	12	85	115
abCd	12	08	85	105
abcD	12	08	73	93
abcd	14	29	60	103

Note. Uppercase letters denote a symptom's presence; lowercase letters denote the symptom's absence.

frequency of occurrence in the training sequence, given a particular symptom pattern, may differ somewhat from the expected value. In this and the subsequent experiments, the correlation between judged probability and both of these normative benchmarks will be reported, computed separately for each participant. In Experiment 1, the mean correlation between judged probability and the Bayesian expected values is 0.71 ($SD = 0.14$, $Mdn = 0.73$), while the mean correlation between judged probability and the actuarial values is 0.68 ($SD = 0.13$, $Mdn = 0.70$). This suggests that participants' judgments were fairly sensitive to the probabilistic category structure.

Table 1 also indicates that the degree of subadditivity, reflected by the total probability T assigned to the three flu strains, varies considerably over the 16 patterns. Two factors considered in the introduction, cue conflict and cue frequency, were assessed for their ability to account for the variance in T using multiple regression analysis. The number of flu strains implicated by present symptoms in the symptom pattern is taken as an index of cue conflict. By this measure, for example, the degree of cue conflict associated with the symptom pattern ABcd is 2, because the presence of Symptom A implicates Flu Strain 1 and the presence of Symptom B implicates Flu Strain 2. The value of the cue conflict variable ranges from 0 (for pattern abcd) to 3 (e.g., pattern ABCd). The cue frequency factor was represented in the analysis by the average relative frequency (over the course of the training phase of the experiment) of the present cues in the symptom pattern. So, for example, the cue frequency value associated with symptom pattern ABcd, given that the relative frequency of Symptoms A and B is 56% and 46% respectively, is 51%. The value of the cue frequency variable ranges from 0% (for pattern abcd, which had no present symptoms) to 56% (e.g., pattern ABCd). The dependent variable is the total probability T assigned by a participant for a particular symptom pattern.

Table 2
Results From Regression Analysis of Total Probability T *(in Percent) in Experiments 1–5, With Cue Conflict, Cue Frequency, and Cue Redundancy (Experiments 3–5 Only) as Predictors*

Experiment and predictor	B	SE of B	β	% individual $B > 0$
Experiment 1				
Conflict	7.51**	2.74	0.18	81
Frequency	0.24	0.19	0.09	63
Experiment 2[a]				
Conflict	11.88**	2.62	0.22	77
Frequency	−0.21	0.16	−0.06	42
Experiment 3				
Conflict	8.46**	1.80	0.17	88
Frequency	3.48	2.21	0.06	69
Redundancy	7.07**	2.21	0.12	63
Experiment 4				
Conflict	6.09**	1.60	0.13	75
Frequency	1.86	1.94	0.03	58
Redundancy	1.39	1.95	0.02	54
Experiment 5[b]				
Conflict	15.37**	2.24	0.26	89
Frequency	1.48	2.74	0.02	63
Redundancy	6.94*	2.74	0.10	63

Note. Final column indicates proportion of participants for whom $B > 0$ in individual-level regression analyses. Cue conflict is represented by the number of flu strains implicated by symptoms present in the symptom pattern. Cue frequency is represented by the average symptom frequency in percent computed over symptoms present in the pattern in Experiments 1–2 and by the dummy-coded presence of Symptom D versus Symptom E in Experiments 3–5. Cue redundancy is represented by the number of symptoms present in the pattern minus the number of flu strains implicated by pattern. Effects of experimental variables were estimated simultaneously with a subject variable to control for individual differences in general overestimation or underestimation.

[a]Experiment 2 analysis is based on mean probability judgments for each participant calculated over the last 180 training trials. [b]The similarity-rating measure acting as the dependent variable in Experiment 5 is multiplied by a factor of 10 for purposes of scale comparability across experiments.

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

Table 3
Mean Value of T *(in Percent) in Experiments 1–5 by Level of Cue Conflict*

Experiment	Level of cue conflict[a] 0	1	2	3
1	102	107	129	123
2[b]	114	117	130	139
3	132	135	148	153
4	116	127	135	132
5[c]	127	135	156	167

[a]Level of cue conflict represents the number of flu strains implicated by symptoms present in the symptom pattern. [b]Experiment 2 values are based on mean probability judgments for each participant calculated over the last 180 training trials. [c]The similarity-rating measure acting as the dependent variable in Experiment 5 is multiplied by a factor of 10 for purposes of scale comparability across experiments.

Table 2 presents the results of the regression analysis in this and subsequent experiments reported in the article. In all analyses, the main experimental independent variables were entered simultaneously with a subject variable to control for individual differences in general overestimation or underestimation. The regression model used for each experiment includes only main effect terms for each independent variable, as results across the five experiments generally failed to show any significant advantage for more complex models including cross-product (i.e., interaction) terms. (See Experiment 5 for the one exception.)

Results of the regression analysis for Experiment 1 demonstrate a significant effect of cue conflict. Table 3 lists the mean value of T by level of conflict in this and subsequent experiments reported in this article. The table indeed shows that T generally increased with level of cue conflict, though the relationship is not completely monotonic. The source of this nonmonotonicity is not clear, and it reappears in only one of the four subsequent experiments.

Cue frequency did not have a comparable effect on the value of T. Recall that Symptom A was more frequent overall than B in the training sequence, B was more frequent than C, and so on. Symptom patterns with a higher

frequency in the training sequence were anticipated to be associated with greater subadditivity. Indeed, inspection of only those patterns with a single present symptom (i.e., Abcd, aBcd, abCd, abcD) reveals a trend in this direction: The mean total probabilities for these four patterns are 116%, 109%, 105%, and 93%, respectively, showing that subadditivity did tend to increase with cue frequency. A more general effect of cue frequency across all the symptom patterns, however, does not seem to emerge. In fact, the overall effects of cue frequency appear to be fairly negligible in this and subsequent experiments.

Summary. The total probability T assigned to the three flu strains for a given symptom pattern, taken as a measure of subadditivity, systematically exceeded 1, contrary to the rules of probability theory but consistent with the predictions of support theory. The value of T varied substantially and systematically from one symptom pattern to the next. Increased cue conflict (but not cue frequency) was associated with enhanced subadditivity: The greater the number of flu strains implicated by symptoms present in the symptom pattern, the greater the value of T. This result supports the findings of previous research (Koehler et al., 1997; Peterson & Pitz, 1988) using a much different task in which judgments are based not on general knowledge but rather on knowledge acquired during the course of the experiment.

Experiment 2

The first experiment revealed substantial subadditivity in probability judgments elicited following learning. It could be argued, however, that had probability judgments been elicited within the learning context, instead of after learning had taken place, the general observation of subadditivity might have been eliminated. Providing feedback immediately after each probability judgment, for example, might draw participants' attention to the fact that their judgments are generally too high, consequently reducing or even eliminating the subadditivity found in the post-learning judgments. This possibility was tested in Experiment 2 by asking participants to make a probability judgment on each training trial, using a training sequence identical to that of Experiment 1. As in the first experiment, the symptom patterns upon which participants based their judgments varied in terms of cue conflict and cue frequency, allowing assessment of the influence of these factors on the total probability T assigned to the three flu strains.

Method

Participants. Participants were 34 prospective psychology undergraduate majors at University College London, who participated as part of a laboratory demonstration. As elaborated below, data from three of these participants were dropped as their learning performance was only marginally better than that expected by chance, leaving a total of 31 participants.

Stimuli and apparatus. The stimuli and the training sequence used were identical to that of Experiment 1. The experiment was conducted using IBM PC-compatible computers, which presented the symptoms and judgments using essentially the same screen layout as in the first experiment. One minor difference was that, in addition to listing present symptoms in uppercase and absent symptoms in lowercase, the present and absent symptoms were also listed in different colors.

Design. Participants received the same training sequence as in Experiment 1, but assigned a probability to a designated flu strain on each trial rather than choosing which of the three flu strains they thought was most likely. The flu strain designated for evaluation on each trial was varied between participants by assigning each participant to one of three target groups. On any given trial, the three target groups each evaluated one of the three possible flu strains so that, across groups, judgments were obtained of the probability of each flu strain on every trial. Which flu strain was designated for a given target group was determined randomly such that participants in each group were assigned each flu strain with approximately equal frequencies across the training sequence. As in the first experiment, the assignment of symptom names to the abstract category structure and the on-screen presentation order of the symptoms were counterbalanced across participants.

Because participants made probability judgments on every trial of Experiment 2, they were not asked to give a final set pattern judgments at the end of the training sequence as was done in the first experiment. Instead, a comparable set of "pattern judgments" was computed by aggregating the probability judgments made over the last 180 training trials for each combination of symptom pattern and designated flu strain. As in the previous experiment, the resulting set of pattern judgments can be used to investigate the influence of cue conflict and cue frequency. Indeed, the cue conflict and cue frequency values associated with each symptom pattern are identical to that of Experiment 1, because both experiments shared a common training sequence.

Procedure. Instructions regarding the general nature of the medical judgment task were similar to those given for Experiment 1. The major difference is that, in this experiment, participants were instructed to give a probability judgment on every training trial. It was explained that one of the three flu strains would be selected arbitrarily on each trial as the designated outcome for judgment. Because the probability judgments were obtained during learning as individual patients were presented for assessment, the judgments were given a probabilistic interpretation (i.e., the probability that the patient in question has the designated flu strain) rather than a frequentistic interpretation as was given for the judgments obtained following learning in Experiment 1.

Because each participant would have to give 240 probability judgments during the training sequence, a probability judgment scale was provided to allow participants to make their judgments more quickly and easily. So, instead of typing in a number from 0 to 100 on the keyboard as in Experiment 1, participants in Experiment 2 were provided with a scale running from 0% to 100% in increments of 10%. This scale appeared on the screen with a square drawn around the 0% value. Participants moved the square up and down the scale using the left- and right-arrow keys on the keyboard, and pressed enter when the box was on the probability value they wanted. Note that any anchoring effect resulting from this elicitation process would be expected to introduce a bias toward $T < 1$, contrary to the predictions of support theory.

Results and Discussion

Learning performance. Individual learning performance is examined first. In contrast to Experiment 1, in which there was a simple measure of learning (i.e., percent correct), in this second experiment a more complicated analysis is necessary because participants judged the prob-

ability of a designated flu strain rather than chose the flu strain they thought was most likely. Perhaps the simplest measure is what might be called a *probabilistic hit rate.* If the designated event (flu strain) being judged by the participant actually occurs, then the participant receives a score of p for that trial, where p is the judged probability of the event. If the designated event does not occur, then the participant receives a score of $1 - p$ for that trial. The total score over the full set of learning trials has a maximum of 240, which is achieved only if the participant performs perfectly, that is, assigns a probability of 1 to each event that subsequently occurs and a probability of 0 to each event that subsequently does not occur. Dividing this measure by 240 yields a measure akin to the percent correct measure of Experiment 1.

This measure has the drawback that its value given chance performance (i.e., in the absence of any learning) depends on the participant's response distribution, that is, the frequency with which the participant uses each of the 11 probability categories. To adjust for this, a corrected performance score was computed for each participant by first calculating the expected probabilistic hit-rate associated with chance performance given that participant's response distribution and then subtracting the resulting value from the participant's actual score to obtain a measure of performance above chance. It can be shown that this corrected performance measure is equivalent to an analogously corrected squared error or Brier score measure (Brier, 1950).

All participants performed better than would be expected by chance guessing—that is, all had positive corrected performance measures. For 3 of the 34 participants, however, performance was only marginally better than chance. These participants (one from each of the 3 target groups) were clear outliers, with corrected performance measures that were more than 1.8 standard deviations below the mean of the rest of the group, and thus were dropped from subsequent analysis. For the remaining 31 participants, the mean corrected performance value was 26.3, corresponding to an average probabilistic hit rate of 63.6%. The best-performing participant had a corrected performance value of 54.0, the worst a value of 12.5.

As in the previous experiment, mean learning performance was examined for the four sequential sets of 60-trial blocks. The mean corrected performance value (computed separately for each participant and then averaged) was 0.8, 8.3, 8.7, and 8.6 for Blocks 1, 2, 3, and 4, respectively. This analysis suggests that, as in the first experiment, learning was at or near asymptote after the first 60 trials or so. A simple percent correct measure comparable to that of Experiment 1 was computed at the group level by selecting as the "chosen" flu strain on each trial that flu strain receiving the highest mean probability judgment and then computing the percentage of trials on which the flu strain so chosen was correct. For Trial Blocks 1 to 4, the percent correct classifications using this measure was 37%, 71%, 67%, and 65%, respectively. These values are comparable to the corresponding figures for Experiment 1 and reinforce the conclusion that learning was at asymptote after approximately 60 trials.

Pattern judgments. In contrast to Experiment 1, in which a final set of pattern judgments was elicited after learning, in Experiment 2 the pattern judgments were obtained during the training sequence itself. As the above analysis suggests that learning was at or near asymptote by Trial 60, the pattern judgments were obtained by averaging over trials 61–240. Table 4 lists the mean judgment assigned to each flu strain, and their total, for each of the 16 possible symptom patterns. Note that some patterns occurred more frequently than others during the training sequence, as determined by the probabilistic category structure. As a result, the mean judgments are based on different numbers of observations for the different patterns.

As in the previous experiment, participants' judgments corresponded quite closely to the normative values. The single-symptom patterns, for example, yielded high probabilities for the associated flu strains and low probabilities for the others. The mean correlation between judged probability and the Bayesian expected values is 0.61 ($SD = 0.15$, $Mdn = 0.67$), while the mean correlation between judged probability and the actuarial values is 0.58 ($SD = 0.15$, $Mdn = 0.63$). These values are somewhat lower than those obtained in the first experiment. The correlation between the set of mean pattern judgments obtained in Experiments 1 and 2 is 0.95.

More importantly for present purposes, the probability judgments were subadditive for all 16 symptom patterns, as can be seen in Table 4. The (unweighted) mean total probability T assigned to the three possible flu strains is 124%, which is slightly greater than the comparable value of 120% for the pattern judgments of Experiment 1. Participants' judgments were consistently subadditive, then, even when feedback immediately followed every probability judgment, and even though the probability scale started with

Table 4
Average Judged (Percent) Probability of Each Flu Strain for the 16 Possible Symptom Patterns, Along With Their Total T, *Computed Over the Last 180 Learning Trials of Experiment 2*

Pattern	Flu 1	Flu 2	Flu 3	Total T	n
ABCD	34	29	41	104	1
ABCd	38	52	52	143	8
ABcD	38	44	50	132	4
ABcd	57	56	17	130	27
AbCD	39	24	75	138	6
AbCd	62	24	47	133	16
AbcD	55	31	44	130	10
Abcd	82	18	11	111	32
aBCD	19	31	70	120	4
aBCd	16	60	56	132	9
aBcD	12	63	52	127	2
aBcd	17	79	18	114	24
abCD	21	29	77	128	7
abCd	18	24	73	116	14
abcD	20	29	72	120	8
abcd	18	28	66	112	8

Note. Uppercase letters denote a symptom's presence; lowercase letters denote the symptom's absence. The number of times the pattern appeared in the 180 learning trials is designated by n.

an anchor of 0%. Indeed, comparison with the pattern judgments of Experiment 1 suggests that the change in experimental procedure did nothing at all to reduce the degree of subadditivity in the pattern judgments.

Given the observation of general subadditivity, the roles of cue conflict and cue frequency can be examined (see Table 2). As in the previous experiment, cue conflict had a significant effect, such that T increased with level of conflict (see Table 3 for means). Once again, cue frequency had no significant effect on T.

Summary. Probability judgments elicited during the training phase in Experiment 2 exhibited the same general pattern as probability judgments elicited following the training phase in Experiment 1. These judgments were systematically subadditive, as indicated by $T > 1$, with the degree of observed subadditivity increasing with cue conflict. Once again, cue frequency had no significant effect on the value of T.

Experiment 3

A third experiment was conducted to assess the generalizability of the effects of cue conflict and cue frequency, by using a different category structure than that used in the first two experiments. The new category structure introduced a fifth symptom, effectively doubling the number of distinct symptom patterns and thus providing a larger sample of judgments for testing effects of the experimental variables.

The new category structure was also intended to completely separate testing of the effects of cue conflict and cue frequency. In the resulting design, cue conflict can be tested using cues that are equated in terms of frequency, and cue frequency can be tested using cues that are completely nondiagnostic with respect to the outcome variable. Cue frequency was varied over a wider range than in the previous experiments, providing a stronger test of the hypothesis that subadditivity increases with cue frequency. In addition, the new design allows investigation of a third factor, cue redundancy, that might also influence the degree of subadditivity associated with a particular symptom pattern.

Finally, the third experiment also eliminated a potential problem with the first two experiments, namely that all participants received an identical training sequence. To ensure that the results of Experiments 1 and 2 are not in some way an idiosyncratic result of the particular training sequence provided to all participants (see Lewandowsky, 1995), in this experiment a different randomly ordered training sequence was created for each participant in the experiment.

Method

Participants. Participants were 16 undergraduates at the University of Waterloo, who participated in exchange for credit in their introductory psychology course. Data from two additional participants were dropped: one whose learning performance was not greater than that expected by chance, and one who reported to the experimenter that she had failed to complete the judgment task as instructed.

Stimuli and apparatus. The experiment was conducted using IBM PC-compatible computers, which presented the symptoms using the same screen layout as in Experiment 2. The only change was that participants made a choice decision (as in Experiment 1) by moving a box to select the flu strain they diagnosed and pressing the return key, rather than a probability judgment (as in Experiment 2) on each of the training trials.

Design. The major difference from the first two experiments is that a new category structure was introduced. Participants were presented with information regarding five symptoms, rather than four as in the previous experiments. The training sequence again consisted of 240 trials. This sequence was constructed by first randomly choosing one of the three flu strains (with equal probabilities) and then choosing the five symptoms (independently) with the appropriate probabilities for that flu strain.

Symptoms A, B, and C were equally diagnostic and were associated with Flu Strains 1, 2, and 3, respectively. The likelihood of the presence of the symptom associated with a flu strain (e.g., of Symptom A given Flu Strain 1) increased to 75% in the presence of that flu strain and decreased to 25% in its absence. Consequently, the conditional probability of a flu strain given the presence of its associated symptom (e.g., of Flu Strain 1 given Symptom A) was 60%, with the remaining two flu strains having a probability of 20% each. Given the absence of a symptom, by contrast, the conditional probability of its associated flu strain dropped to 14%, while the probability of the two alternative flu strains increased to 43% each. Because each symptom implicates a different flu strain, Symptoms A–C provide the basis for testing the influence of level of cue conflict, defined as the number of these symptoms present in the symptom pattern.

Symptoms D and E were nondiagnostic and differed only in terms of their overall frequency of occurrence in the training sequence. Regardless of the patient's flu strain, Symptom D was present with a probability of 75%, while Symptom E was present with a probability of 25%. Note that this represents a greater difference in cue frequency than that investigated in the first two experiments, allowing a stronger test of cue frequency's influence on judged probability. An effect of cue frequency would be demonstrated in this design if those symptom patterns including D but not E were associated with greater subadditivity than those symptom patterns including E but not D.

Because they are nondiagnostic, Symptoms D and E also provide a basis for testing the effect of cue redundancy, that is, the influence of the mere presence of symptoms independent of any diagnostic value they might possess. If cue redundancy is associated with enhanced subadditivity, then symptom patterns in which both D and E are present would be expected to yield greater subadditivity than symptom patterns in which both D and E are absent, with symptom patterns in which only one of the two symptoms are present expected to fall in between.

A set of 240 patient cases was constructed using these probabilities such that the relative frequencies of the symptoms given each of the flu strains was achieved exactly over the training sequence. The order in which the 240 patients were presented in the training sequence was determined randomly for each participant. The actual symptom (e.g., cough) assigned to the five abstract Symptoms A–E was once again counterbalanced over participants, as was the position of the five symptoms in the computer display.

The introduction of a fifth symptom (dizziness) increased the number of pattern judgments to 96, the order of which was determined randomly for each participant. Participants made their judgments—which were given a probabilistic interpretation—using a probability judgment scale running from 0% to 100% in increments of 10% as in the previous experiment.

Procedure. The general procedure and the instructions given to participants were essentially identical to those of Experiment 1. The only procedural difference is that participants entered their choices and probability judgments into the computer, as in Experiment 2, by moving a box via the arrow keys on the keyboard until their desired response was selected and pressing the return key.

Results and Discussion

Learning performance. Over participants, average accuracy across the 240 training trials was 47%. The most accurate participant achieved 56% correct, and the least accurate achieved 42% correct. Generally speaking, participants were less accurate in the training phase of this experiment than they were in the previous two. Such a result would be expected given the changes in the category structure introduced in this experiment: Participants had to consider five symptoms (rather than four as in the previous experiments), only three of which were diagnostic. All participants included in the sample achieved significantly above-chance accuracy.

The theoretical maximum percent correct that could be achieved in this task was assessed in terms of the accuracy achieved by a maximizing strategy, as was done in Experiment 1. This strategy yielded an accuracy level of 67.5% across the 240 trials. This analysis suggests—consistent with the learning performance results above—that the participants' task was somewhat more difficult than in the previous experiments (for which the maximizing strategy yielded an accuracy level of 72%).

Improvement over trials was relatively modest compared to the previous experiments. Once again, average percent correct (over participants) was computed for four consecutive 60-trial blocks. On the first block, 45% of participants' guesses were correct. For the next three blocks the corresponding figures were 44%, 49%, and 50%, respectively. Participants' performance showed little sign of improvement in the second half of the training sequence, suggesting that by the end of the training phase participants had learned all they could about the probabilistic category structure.

Pattern judgment data. Table 5 presents the mean probability judgments assigned to each flu strain for the 32 possible symptom patterns. Consistent with the predictions of support theory, these probability judgments were clearly subadditive: The total probability T assigned to the three possible flu strains consistently exceeded 100%, with an average total of 142%. Individually, only one of the sixteen participants had an average total (M = 97%) of less than 100%. The degree of subadditivity observed for these judgments appears to be considerably greater than that of the previous experiments, as might be expected if the inclusion of an additional symptom induced a greater sense of conflict or uncertainty.

Once again, participants' probability judgments were generally in accord with information encountered in the training sequence. Thus, for patterns that have only a single symptom present, participants gave high probabilities to the appropriate flu strain (e.g., Flu Strain 1 was assigned a mean probability of 74% given the pattern Abcde). The mean correlation between judged probability and the Bayesian expected values is 0.44 (SD = 0.17, Mdn = 0.42), while the mean correlation between judged probability and the actuarial values is 0.38 (SD = 0.12, Mdn = 0.35).

As in the previous two experiments, cue frequency and cue conflict were tested as predictors of the total probability T assigned for a particular symptom pattern. The new category structure allows a stronger test of cue frequency by a direct contrast between symptom patterns that include the more frequent Symptom D but not the less frequent Symptom E (coded +1) and symptom patterns that include the less frequent Symptom E but not the more frequent Symptom D (coded −1), with the remaining symptom patterns coded 0.

The new category structure also allows testing of an additional factor, referred to as cue redundancy. This factor reflects the effect of the presence of additional present symptoms, holding constant the number of flu strains implicated by the symptom pattern. As an illustration, consider the symptom patterns AbCde, AbCDe, and AbCDE. All three have the same cue conflict value (2), because in all three the present symptoms A and C taken together implicate two flu strains (1 and 3, respectively). The three patterns vary, however, in the total number of present

Table 5
Average Judged (Percent) Probability of Each Flu Strain for the 32 Possible Symptom Patterns, Along With Their Total T, *for Experiment 3*

Pattern	Flu 1	Flu 2	Flu 3	Total T
ABCDE	38	54	53	145
ABCDe	50	56	64	170
ABCdE	39	57	49	145
ABCde	44	49	59	152
ABcDE	52	61	42	155
ABcDe	58	61	29	148
ABcdE	55	55	34	144
ABcde	61	44	28	133
AbCDE	54	42	60	156
AbCDe	50	41	62	153
AbCdE	53	41	49	143
AbCde	42	39	56	137
AbcDE	63	44	35	142
AbcDe	78	29	21	128
AbcdE	56	39	28	123
Abcde	74	35	11	120
aBCDE	36	60	75	171
aBCDe	23	54	74	151
aBCdE	29	49	68	146
aBCde	18	44	71	133
aBcDE	39	68	39	146
aBcDe	46	71	35	152
aBcdE	39	61	36	136
aBcde	31	66	31	128
abCDE	26	38	70	134
abCDe	27	38	81	146
abCdE	28	45	74	147
abCde	9	32	77	118
abcDE	44	49	40	133
abcDe	45	51	26	122
abcdE	39	41	47	127
abcde	59	21	66	146

Note. Uppercase letters denote a symptom's presence; lowercase letters denote the symptom's absence.

symptoms they include (by virtue of the presence or absence of the nondiagnostic Symptoms D and E). To test whether this factor also contributes to the total probability T, a cue redundancy variable was added to the regression analyses, defined as the total number of present symptoms in the symptom pattern minus the number of implicated flu strains (i.e., total present symptoms − cue conflict index). In the present experimental design, cue conflict, cue frequency, and cue redundancy are completely uncorrelated variables. Table 2 shows the results of this analysis.

Once again, the total probability T increased significantly with the degree of cue conflict (see Table 3 for means). Cue redundancy was also positively associated with the total probability. That is, the additional presence of Symptoms D or E in the symptom pattern generally increased T. Table 6 lists the mean value of T by level of cue redundancy in this and subsequent experiments reported in this article. As in the previous experiments, cue frequency had no significant effect.

Summary. Using a new category structure involving five symptoms, probability judgments based on symptom patterns exhibited an even greater degree of subadditivity than that observed in the first two experiments. Both cue conflict and cue redundancy were found to influence the degree of subadditivity associated with a particular symptom pattern: The greater the number of flu strains implicated by symptoms present in the symptom pattern, and the greater the total number of present symptoms, the greater was the value of T. Despite being varied over a greater range of values, cue frequency failed to have any substantial influence.

Experiment 4

Psychologically, judgments of probability and of frequency can evoke different responses, even under conditions in which standard normative analyses would regard them as equivalent (Gigerenzer & Hoffrage, 1995; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Griffin & Buehler, 1999; Griffin & Tversky, 1992; Kahneman & Tversky, 1979, 1982; Reeves & Lockhart, 1993; Teigen, 1974b). In particular, it has been noted that a frequentistic formulation often produces more extensional judgments, that is, judgments that are more consistent with rules of set inclusion. Tversky and Kahneman (1983), for example, found that participants were less likely to make conjunction errors (i.e., estimating the likelihood of a conjunction of events to be greater than one of the constituent events of the conjunction) when the judgment was requested in the form of a relative frequency rather than a probability.

A natural question, then, is whether judgments of frequency (or relative frequency) will exhibit subadditivity as described by support theory. Tversky and Koehler (1994; also Koehler et al., 1997; cf. Teigen, 1974b) observed that, like probability judgments, relative frequency judgments do exhibit subadditivity. Rottenstreich and Tversky (1997) further demonstrated that judgments of absolute frequency as well as relative frequency are also systematically subadditive. Consistent with the results reported by Tversky and Kahneman (1983), however, the degree of subadditivity observed in judgments of frequency and relative frequency does appear to be less pronounced than that found in probability judgments (Teigen, 1974b; Tversky & Koehler, 1994). Tversky and Koehler suggest that a frequentistic formulation prompts the judge consider a broader range of specific alternatives to the focal hypothesis, thus "unpacking" the implicit disjunction to a greater extent than is the case in judgments of probability. Under this interpretation, support theory predicts that the subadditivity associated with probability judgments should be more pronounced than that associated with comparable frequency judgments.

Recently, a number of researchers (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996; Gigerenzer et al., 1991; Gigerenzer & Hoffrage, 1995) have interpreted the observation that a frequentistic formulation induces more extensional or otherwise normatively appropriate judgments as evidence that the mind has been designed by natural selection to process frequencies rather than probabilities. Gigerenzer and Hoffrage (1995), for example, refer to a modular "cognitive algorithm" in the mind designed for the processing of frequencies: "We assume that as humans evolved, the 'natural' format was *frequencies* as actually experienced in a series of events, rather than probabilities or percentages" (p. 686). They conclude from their studies that "frequency formats made many participants' inferences strictly conform (in terms of outcome and process) to Bayes' theorem without any teaching or instruction" (p. 698). Very similar conclusions were reached by Cosmides and Tooby (1996), who offer an even more explicit evolutionary interpretation, arguing that humans "evolved mechanisms that took frequencies as input, maintained such information as frequentistic representations, and used these frequentistic representations as a database for effective inductive reasoning" (p. 17).

Gigerenzer and Hoffrage (1995, p. 698) noted that their findings may not generalize to more complex problems involving multiple evidential cues and hypotheses, where the Bayesian solution can require highly complicated calculations. Thus, it is an open question whether the systematic subadditive bias observed in the present studies will be reduced or eliminated when the requested judgment is given a frequentistic formulation. While Experiment 1 did provide a frequentistic interpretation for the requested judgments, participants were also provided with an alternative probabi-

Table 6
Mean Value of T *(in Percent) in Experiments 3–5 by Level of Cue Redundancy*

Experiment	Level of cue redundancy[a]		
	0	1	2
3	134	143	148
4	128	129	131
5[b]	139	146	153

[a]Level of cue redundancy represents the total number of present symptoms in the symptom pattern minus the number of implicated flu strains. [b]The similarity-rating measure acting as the dependent variable in Experiment 5 is multiplied by a factor of 10 for purposes of scale comparability across experiments.

listic interpretation and the judgment was made in the form of a relative frequency (in percent). A stronger test requires elicitation of absolute frequency judgments. This is particularly important given that the cognitive algorithm for processing frequency information described by Gigerenzer and Hoffrage (1995, pp. 688–689) is claimed to apply only to absolute frequency, and not relative frequency. To this end, Experiment 4 was conducted in a manner identical to Experiment 3 in all respects except that judgments of absolute frequency rather than probability were elicited. This design provides a test of the general observation of subadditivity, as indicated by $T > 1$, and also tests of the same experimental variables examined in the previous experiment, namely cue conflict, cue frequency, and cue redundancy.

There are actually two aspects to the argument that the mind has a cognitive algorithm designed for processing frequencies. The first, as emphasized above, focuses on the module's output: Elicitation of frequency judgments is predicted to produce better judgments than elicitation of probability judgments. The second aspect concerns input: Such an algorithm should be particularly useful in the analysis of evidence acquired in the form of a series of discrete cases or events. As Cosmides and Tooby (1996) put it, "our hominid ancestors were immersed in a rich flow of observable frequencies that could be used to improve decision-making, given procedures that could take advantage of them. So if we have adaptations for inductive reasoning, they should take frequency information as input" (p. 16; see also Brase et al., 1998, pp. 5–6). Kleiter (1994) has pointed out the potential computational advantages of acquiring and representing such information in the form of frequency counts, under a scheme he refers to as *natural sampling*.

For optimal judgment, then, according to this account, any information provided to participants as a basis for judgment should be presented in the form of a set of single cases, for which running frequency counts can be established. The present experiments employ precisely this kind of design. By modifying the experimental design so that absolute frequency judgments are elicited, on this account, conditions should be optimal for obtaining judgments that are free from any systematic bias. Experiment 4, then, provides a strong test of the hypothesis that there exists a mental module which, given frequentistic information as input, will produce unbiased frequentistic assessments as output in multicue judgment. Support theory, by contrast, implies that such judgments will still exhibit systematic subadditivity, though—as elaborated above—such subadditivity would be expected to be less pronounced than that found in the probability judgments of Experiment 3.

Method

Participants. Participants were 24 undergraduates at the University of Waterloo, who participated in exchange for credit in their introductory psychology course. Data from 8 additional participants were dropped whose learning performance was not significantly greater than that expected by chance guessing.

Stimuli and apparatus. The experiment was conducted using IBM PC-compatible computers, which presented the symptoms using the same screen layout as in Experiments 2 and 3. Training trials proceeded as in Experiment 3, and the screen layout for subsequent judgments was virtually identical as well except for minor changes, as described below, required for the change in judgment format.

Design. The same set of 240 patient cases was used as in Experiment 3. The order in which the 240 patients were presented in the training sequence was determined randomly for each participant. The actual symptom (e.g., cough) assigned to the five abstract Symptoms A–E was again counterbalanced over participants, as was the position of the five symptoms in the display.

As in Experiment 3, there were 32 symptom patterns, which when crossed with the three flu strains, produced a full set of 96 possible pattern judgments. Once again, the full set was elicited from each subject in a randomized order. The level of cue conflict, cue frequency, and cue redundancy varied across the 32 symptom patterns exactly as in the previous experiment, thus allowing analogous tests of the effects of these variables on the degree of subadditivity associated with each symptom pattern.

When given a frequentistic formulation, each pattern judgment has two components: the estimated frequency of patients with the designated symptom pattern, f(pattern), and the estimated frequency of such patients who have the designated flu strain, f(pattern & flu strain). On each of the 96 judgment trials, the f(pattern) estimate and the f(pattern & flu strain) estimate were elicited sequentially. For each symptom pattern, this design yields three f(pattern & flu strain) estimates paired with three f(pattern) estimates; the three f(pattern) estimates associated with a given symptom pattern would be expected to differ only due to unreliability of the estimates. Dividing f(pattern & flu strain) by its paired f(pattern) value produces a measure comparable to the pattern judgments, p(flu strain|pattern), elicited in the previous experiments.

Procedure. Instructions and procedure for the 240 training trials were identical to that of Experiment 3: Participants made a diagnosis choice on each trial, followed by outcome feedback.

On each of the 96 judgment trials, participants first estimated f(pattern) and then were presented with a designated flu strain and asked to estimate f(pattern & flu strain). Participants were instructed as follows:

> In the first part of the experiment, you encountered a total of 240 patients in the process of learning which symptoms tend to go with which flu strains. In the next part of the experiment, you will be asked to estimate the number of patients you encountered out of the total 240 who had a certain pattern of symptoms. That is, you will be presented with a set of symptoms and asked to estimate the number of patients in the original group of 240 you saw who had that EXACT set of symptoms.

The symptom pattern appeared at the top of the screen, and participants were prompted, "Of the 240 patients you encountered in Part 1 of the study, please estimate the number of patients with the exact set of symptoms above." Participants entered their estimates using the numbers on the keyboard.

After entering their f(pattern) estimate, a flu strain was designated and participants were asked to estimate f(pattern & flu strain). Instructions regarding these estimates were as follows:

> Following your estimate of the TOTAL number of patients with the listed set of symptoms, you will then be asked to estimate the number of such patients you encountered who had a designated flu strain. That is, you will be asked to estimate the number of patients you encountered in Part 1 of

> the study who had the exact set of symptoms listed AND who had the designated flu strain. The top part of the display will list the symptom pattern (set of symptoms). Your estimate of the total number of patients you encountered with that exact set of symptoms will be provided. Your task is to estimate, out of the estimated total number of patients with that exact set of symptoms, how many had the designated flu strain.

As in the previous experiments, participants were informed that the target flu strain would be selected arbitrarily on each trial and should not be treated as a suggestion that the flu strain in question is particularly likely (or unlikely) given the symptom pattern in question. For each f(pattern & flu strain) estimate, subjects were reminded of their f(pattern) estimate (e.g., "You estimated that the number of patients encountered in the first part of the study with this exact set of symptoms was 10") and then were asked for their f(pattern & flu strain) estimate (e.g., "How many of the 10 patients with this set of symptoms had Flu Strain 2? __ out of the 10 patients with this set of symptoms"). Participants' estimates were constrained to be between 0 and their f(pattern) estimate; if they entered a value outside this range, they were reminded of this constraint and asked to enter a revised value.

The dual task of estimating f(pattern) and f(pattern & flu strain) was summarized as follows in the instructions using an example estimate:

> So you will have two tasks on each trial. First, estimate the number of patients you encountered in the first part of the experiment with the set of symptoms listed. Suppose, for example, that your estimate is that you had encountered 20 patients with the particular set of symptoms listed. Second, you will be asked how many of those 20 patients had a designated flu strain (for example, Flu Strain 2). Obviously, your estimate should be less than or equal to the total number of patients you estimated to have the listed set of symptoms (in this example, your estimate should be less than or equal to 20).

The order in which the 96 pairs of estimates were made was randomized for each subject.

Results and Discussion

Learning performance. Over participants, average accuracy across the 240 training trials was 50% (SD = 6.2%). The most accurate participant achieved 63% correct, and the least accurate achieved 38% correct. All participants included in the sample achieved significantly above-chance accuracy. Note, however, that a somewhat larger proportion of the original participants (n = 8 out of the original sample of 32) were dropped from the sample due to failure to achieve significantly above-chance accuracy. There seemed to be more variance in learning performance in this study than was found in Experiment 3; the reason for this is not clear given that the experimental procedure was identical in both studies through the training phase.

Once again, average percent correct (over participants) was computed for four consecutive 60-trial blocks. On the first block, 45% of participants' guesses were correct. For the next three blocks, the corresponding figures were 50%, 52%, and 54%, respectively.

Pattern judgment data. For purposes of comparability to the results of Experiment 3, the "judged" probability of the designated flu strain given a particular symptom pattern was derived by dividing the value f(pattern & flu strain) by the paired value f(pattern) obtained on each pattern judgment trial. One minor complication in this procedure arises in the rare case in which the participant gives an f(pattern) estimate of zero, in which case the probability calculation is undefined. There were only 47 such cases out of a total of 2,304 judgments, representing about 2% of the data. For purposes of analysis below, these judgments were coded as missing values.

Table 7 presents the mean derived probability judgments assigned to each flu strain for the 32 possible symptom patterns. Consistent with the predictions of support theory, the judgments were clearly subadditive: The total derived probability T assigned to the three possible flu strains consistently exceeded 100%, with an average total of 129% (participant SD = 18.7%). (Recoding undefined values arising from f[pattern] = 0 estimates by setting them to zero, instead of treating them as missing data, has a negligible effect on the total average probability, resulting in a mean of 126%.) Individually, every participant in the sample had an average value of T greater than 100%; the lowest value was 101% and the highest was 167%, with a median value of 127.5%. The observation of total probabilities greater than 100% in this experiment does not appear to be an artifact of

Table 7
Average Derived Probability Assigned Each Flu Strain for the 32 Possible Symptom Patterns, Along With Their Total T, for Experiment 4

Pattern	Flu 1	Flu 2	Flu 3	Total T
ABCDE	26	33	52	111
ABCDe	41	52	46	139
ABCdE	46	43	47	136
ABCde	44	42	52	138
ABcDE	48	65	22	135
ABcDe	56	61	22	139
ABcdE	49	65	23	137
ABcde	56	54	19	129
AbCDE	60	42	51	153
AbCDe	48	26	65	139
AbCdE	51	25	57	133
AbCde	47	18	62	127
AbcDE	70	40	20	130
AbcDe	64	33	30	127
AbcdE	73	27	15	115
Abcde	80	26	15	121
aBCDE	26	56	50	132
aBCDe	12	54	69	135
aBCdE	23	50	54	127
aBCde	20	43	70	133
aBcDE	33	71	28	132
aBcDe	26	80	20	126
aBcdE	31	69	22	122
aBcde	29	87	28	144
abCDE	25	31	75	131
abCDe	15	27	79	121
abCdE	27	27	75	129
abCde	17	18	86	121
abcDE	43	43	33	119
abcDe	42	47	31	120
abcdE	45	33	42	120
abcde	40	23	42	105

Note. Uppercase letters denote a symptom's presence; lowercase letters denote the symptom's absence.

range restriction on the judgment scale imposed by small f(pattern) estimates: If only judgments for which f(pattern) > 10 are examined, the average total probability ($M = 132\%$) is, if anything, slightly higher.

Elicitation of judgments of absolute frequency, then, is not sufficient to eliminate the systematic subadditivity observed in previous experiments. Instead, consistent with the predictions of support theory on the assumption that a frequentistic formulation invokes greater spontaneous unpacking than does a probabilistic formulation (see Tversky & Koehler, 1994, pp. 550–552), frequency judgments exhibit subadditivity that is substantial but less pronounced than that found in judgments of probability. Indeed, while the mean total probability of 129% observed in this experiment is significantly greater than 100%, $t(23) = 7.65$, $p < .001$, it is also significantly lower than the mean total probability of 142% observed in Experiment 3 by a one-tailed test, $t(38) = 1.84$, $p < .05$. Thus, a frequentistic formulation reduces but does not eliminate the systematic subadditivity observed in the previous experiments.

As was the case with the probability judgments in the previous experiments, participants' frequency judgments were quite sensitive to the probabilistic category structure. The correlation between the estimated probability values (derived from the frequency judgments) and the Bayesian expected values is 0.54 ($SD = 0.17$, $Mdn = 0.57$). This correlation is marginally greater than the corresponding correlation found between the pattern judgments of Experiment 3 and the same expected values, $t(38) = 1.73$, $p < .10$, suggesting that the frequentistic formulation of the judgments in Experiment 4 led to somewhat improved accuracy in the correlational sense. The correlation between the estimated probability judgments and the actuarial values is 0.44 ($SD = 0.14$, $Mdn = 0.47$). This value is also greater than that found in Experiment 3, though the difference is not statistically significant, $t(38) = 1.45$. The correlation between the set of mean pattern judgments in Experiment 4 and those from Experiment 3 is 0.87.

Subadditivity in this experiment can be measured by the total probability assigned to the three flu strains, as above; alternatively, it can be assessed by direct examination of the frequency judgments themselves. Support theory implies that increasingly refined partitions of a sample space will produce increasing subadditivity. In this case, such subadditivity can be assessed by the total frequency count assigned across the partition. Normatively, given the training set of 240 patients, any partition of that set of patients should produce a total frequency count of 240, regardless of the particular partition that is employed. The data do not exhibit such a pattern. Instead, consistent with support theory, greater total frequencies are associated with finer partitions. When participants are asked to assess the frequencies of occurrence of the 32 possible symptom patterns in the training sequence, the total mean frequency count assigned across this partition is 465, almost twice the normative value of 240. When participants are asked to assign frequencies to an even finer partition consisting of the 96 possible symptom-pattern/flu-strain conjunctions, the total mean frequency count assigned to the resulting partition is 606, which is significantly larger than that assigned to the symptom pattern partition, paired $t(31) = 4.40$, $p < .001$. This difference between the two partitions is essentially redundant with the observation of $T > 1$.

As in the previous experiment, the influence of cue conflict, cue frequency, and cue redundancy on the value of T can be examined (see Table 2). Once again, the value of T increased significantly with level of cue conflict (see Table 3 for means; as in Experiment 1, the relationship is not completely monotonic at the highest value of cue conflict). Cue frequency again failed to have a significant effect, though a trend in the direction of T increasing with cue frequency was apparent. Finally, cue redundancy did not have a statistically significant effect in this experiment, though a trend in the direction consistent with the previous experiment was observed (see Table 6 for means).

Summary. Elicitation of judgments of absolute frequency did not eliminate the pattern of subadditive judgments observed in the previous experiments. Instead, as predicted by support theory, judgments of frequency were systematically subadditive but less so than comparable judgments of probability. The degree of subadditivity observed in the frequency judgments, like that found for probability judgments, varied systematically with cue conflict, that is, the number of flu strains implicated by symptoms present in the symptom pattern. The effects of cue redundancy observed in Experiment 3, however, were not replicated in Experiment 4. Cue frequency failed once again to have a significant effect on the degree of observed subadditivity.

Experiment 5

The results of Experiments 1–4 demonstrate considerable subadditivity in probability (and frequency) judgments in a classification-learning task. The observation of general subadditivity, as measured by the total probability T assigned to a set of elementary hypotheses, is consistent with the predictions of support theory. Tversky and Koehler (1994) introduced a discounting factor w as a more refined measure of subadditivity that is indexed to a particular implicit disjunction. In the present context of three elementary hypotheses F_1, F_2, and F_3, representing the three possible flu strains, consider as an example the elementary judgment $P(F_1, \overline{F}_1)$. Letting the discounting factor $w_{\overline{F}_1}$ represent the extent to which the support for hypotheses F_2 and F_3 is discounted by their implicit inclusion in the residual $\overline{F}_1$, the value of the elementary judgment is given by

$$P(F_1, \overline{F}_1) = \frac{s(F_1)}{s(F_1) + w_{\overline{F}_1}[s(F_2) + s(F_3)]}, \qquad (3)$$

where $w_{\overline{F}_1} \leq 1$ according to support theory. The lower the value of the discounting factor, the greater the degree of subadditivity.

Koehler et al. (1997) introduced a particular form for the relationship between the support for the focal hypothesis and the extent to which the resulting residual hypothesis is discounted, called the *linear-discounting model.* According

to this model, the support of the hypotheses included implicitly in the residual $\bar{F}_1$ is discounted by a factor $w_{\bar{F}_1}$ that decreases in a linear fashion as the support for the focal hypothesis increases:

$$w_{\bar{F}_1} = 1 - \beta s(F_1). \quad (4)$$

The model is intended to capture the intuition that when support for the focal hypothesis is high, the corresponding residual hypothesis may be unpacked into its components to a lesser extent, and evidence supporting the individual components of the residual hypothesis may be evaluated less exhaustively, than when support for the focal hypothesis is low. Simply put, when the focal hypothesis appears consistent with the evidence, the judge may be less inclined to assess the implications of the evidence for individual hypotheses included implicitly in the residual than when the focal hypothesis appears less consistent with the evidence. The linear-discounting model gives rise to the enhancement effect, in that increasing the support for a set of hypotheses will produce more discounting and hence enhanced subadditivity.

To fit the model, however, the support for each hypothesis provided by a given body of evidence must be assessed. Tversky and Koehler (1994; also Koehler et al., 1997) have shown that people are able to provide direct assessments of evidential support that can then be used to predict probability judgments obtained from a separate group of participants. A final experiment was conducted to collect the required set of direct support assessments, which would allow fitting of the linear-discounting model to the probability and frequency judgment data of Experiments 3 and 4.

Experiment 5 used the same training sequence and category structure as that used in Experiments 3 and 4, but instead of asking participants for probability or frequency judgments, they were asked to make judgments that could be used to estimate the evidential support for each flu strain provided by each symptom pattern. This was accomplished by asking participants—after the training sequence—to rate the degree to which a particular patient (characterized by a symptom pattern) resembled the prototypical patient suffering from a designated flu strain. It was assumed that the similarity judgments would provide a reasonably good measure of the perceived evidential support for a designated flu strain provided by a particular symptom pattern. The similarity judgments can be used to fit the Koehler et al. (1997) linear-discounting model to the probability judgments collected in Experiment 3 and to the frequency judgments collected in Experiment 4. Furthermore, as in the previous experiments, the influences of cue conflict, cue frequency, and cue redundancy on the judgments can be assessed.

Method

Participants. Participants were 19 undergraduates at the University of Waterloo, who participated in exchange for credit in their introductory psychology course. Data were dropped from one additional participant whose learning performance was only marginally greater than that expected by chance guessing.

Design. The 96 pattern judgment trials were blocked by target flu strain to simplify the judgment task. The order in which patient cases were encountered within a block was determined randomly for each participant on each block. In all other respects, the experimental design was identical to that of Experiments 3 and 4.

Procedure. The training portion of the experiment proceeded exactly as in Experiments 3 and 4. The major change in this experiment was the introduction of a new dependent measure on the pattern judgment trials. On these trials, participants were presented with a target "patient" case. Their task was to assess how similar the target patient was to their prototype of the typical patient suffering from a designated flu strain. Participants were told:

> By now, you have probably developed some understanding of the relationships between various symptoms and the three possible flu strains. Put differently, you probably have a prototype or image of what the typical patient suffering from a particular flu strain looks like, in terms of the set of symptoms you have been studying. In the next part of the experiment you will be presented with a set of patients, just as in the last section. In this case though, you are asked to judge the SIMILARITY of the patient in question to your prototype or image of the typical patient suffering from a designated flu strain, for example, Flu Strain 1. In other words, you will be asked to judge how similar the target patient is to your mental image or prototype of a typical patient suffering from the designated flu strain.

Judgments were made on a 0–10 scale where 0 was labeled "not at all similar" and 10 was labeled "highly similar." Participants were instructed:

> A similarity rating of 0 means the target patient is not at all similar to your prototype or image of the typical patient suffering from the designated flu strain. A similarity rating of 10 indicates that the target patient is highly similar to your prototype or image. Intermediate ratings indicate intermediate degrees of similarity.

Once again, participants entered their judgments by moving a box via the arrow keys on the keyboard until their desired response (0–10) was selected and then pressing the return key. In all other respects, the procedure was identical to that of Experiments 3 and 4.

Results and Discussion

Learning performance. Over participants, average accuracy across the 240 training trials was 48%, which is essentially identical to that achieved by participants in Experiment 3. The most accurate participant achieved 58% correct, and the least accurate achieved 39% correct. All participants included in the sample achieved significantly above-chance accuracy.

Participants' performance showed little sign of improvement in the second half of the training sequence, suggesting that by the end of the training phase participants had learned all they could about the category structure. On the first block of 60 trials, 41% of participants' guesses were correct. For the next three blocks, the corresponding figures were 47%, 51%, and 51%, respectively.

Pattern judgment data. Table 8 presents the mean similarity judgments assigned to each of the 32 possible symptom patterns for each of the three target flu strains. The

Table 8
Average Similarity Ratings Between Each Flu Strain and Each of the 32 Possible Symptom Patterns in Experiment 5

Pattern	Flu 1	Flu 2	Flu 3
ABCDE	4.37	3.79	6.79
ABCDe	4.84	5.74	6.79
ABCdE	4.84	4.89	7.11
ABCde	5.68	5.00	7.00
ABcDE	6.00	6.32	4.95
ABcDe	5.74	5.95	4.37
ABcdE	6.16	5.53	4.32
ABcde	6.95	6.47	3.21
AbCDE	4.89	2.89	7.05
AbCDe	6.05	4.53	6.84
AbCdE	5.00	4.42	6.53
AbCde	5.37	3.26	5.63
AbcDE	7.00	4.89	3.37
AbcDe	7.47	4.21	2.89
AbcdE	6.89	4.47	2.89
Abcde	7.63	4.21	2.89
aBCDE	3.37	5.00	7.11
aBCDe	2.74	4.42	7.32
aBCdE	3.16	5.32	6.74
aBCde	3.32	4.26	6.47
aBcDE	5.00	7.11	3.63
aBcDe	4.21	5.89	2.63
aBcdE	4.42	5.68	3.95
aBcde	2.95	6.84	2.26
abCDE	3.37	4.47	6.53
abCDe	3.05	3.05	5.89
abCdE	2.95	3.05	6.11
abCde	2.32	2.68	5.11
abcDE	5.63	5.47	3.53
abcDe	4.32	5.21	3.68
abcdE	4.42	4.00	2.63
abcde	3.32	3.00	5.58

Note. Uppercase letters denote a symptom's presence; lowercase letters denote the symptom's absence.

judgments appeared to correspond in a reasonable manner to the information presented during the training phase. For example, for patterns that have only a single symptom present, participants gave high ratings to the appropriate flu strain (e.g., the pattern aBcde was assigned a mean similarity rating of 6.84 to Flu Strain 2 but only 2.95 and 2.26, respectively, to Flu Strain 1 and Flu Strain 3). The correlation between the mean similarity ratings and the Bayesian expected values was 0.32 ($SD = 0.28$, $Mdn = 0.27$), while the correlation between the mean ratings and the actuarial values was 0.26 ($SD = 0.23$, $Mdn = 0.30$), showing that participants' similarity ratings were not as strongly related to the normative values as were the probability and frequency judgments elicited in Experiments 3 and 4. Over the set of 96 mean judgments, the judged similarity measure had a correlation of 0.81 with judged probability (from Experiment 3) and 0.80 with judged frequency (from Experiment 4).

Results from the earlier experiments were interpreted by postulation of enhanced evidential support available for each flu strain under high cue conflict. If the similarity judgments provide a measure of support, then we would expect to find an influence of cue conflict comparable to that found in judgments of probability. Indeed, regression analysis of the total of the similarity judgments assigned to the three flu strains given a particular symptom pattern (see Table 2) did indicate a significant effect of cue conflict. As predicted, then, the presence of high cue conflict induced greater perceived support for the hypotheses, as measured by the similarity ratings, than was found under low cue conflict (see Table 3 for means).

Cue redundancy also had a significant influence on the similarity ratings (see Table 6 for means), such that the number of symptoms present in the pattern was associated with greater judgments (in this case, the similarity proxy for evidential support), even when controlling for the number of flu strains implicated by the symptom pattern. Once again, cue frequency was found to have little influence.

In the previous four experiments, more complex regression models including interaction terms failed to account for significantly more variance in T than the simpler model including only main effect terms. Results from Experiment 5 represent the one exception: A model including interaction terms accounted for a small (1.3%) but statistically significant increase in R^2, due to a significant conflict by redundancy interaction ($B = -9.28$, SE of $B = 3.14$, $p < .01$). Inspection of cell means (see Table 8) reveals that this interaction is attributable to the mean total similarity (based on one observation per participant) given symptom pattern ABCDE being lower than would be expected given independent main effects of cue conflict and redundancy.

Model-fitting results. The linear-discounting model introduced by Koehler et al. (1997) was fit to the judgment data of Experiments 3 and 4, using the similarity judgments as a measure of support. Fitting of the model involves estimating the value of two free parameters. First, it was assumed that the similarity ratings were related to the perceived evidential support for a hypothesis via a power transformation with exponent θ of the form

$$s_X(F_n) = sim(F_n, X)^\theta, \quad (5)$$

where $s_X(F_n)$ is the support for flu strain n provided by the symptom pattern X, and $sim(F_n, X)$ is the rated similarity of a patient with symptom pattern X to the prototypical patient suffering from flu strain n. This transformation has been used with reasonable success in previous work (Koehler, 1996; Koehler et al., 1997; Tversky & Koehler, 1994). It can be shown that such a relationship must hold between the support and similarity scales if (a) the two scales are monotonically related, and (b) ratios of values on the two scales are also monotonically related (see Tversky & Koehler, 1994). Second, once the similarity ratings have been transformed to support values, it was assumed that judged probability is related to support as specified by the linear-discounting model described in Equation (4), with β as a free parameter.

This model was used to fit the set of 96 mean similarity ratings from Experiment 5 to the corresponding set of 96 mean probability judgments from Experiment 3 and to the comparable set of 96 mean "derived" probability judgments computed from the frequency judgments in Experiment 4. The two free parameters, θ and β, were estimated separately for each data set using a least-squares fitting procedure.

Table 9
Fits of Linear-Discounting Model, Constant-w Model, and Noncompensatory Model to the Mean Probability Judgment Data of Experiment 3 and the Mean Frequency Judgment Data of Experiment 4

Data set and model (parameter values)	Pattern judgments		Total T	
	RMSE	R^2	RMSE	R^2
Experiment 3 (prob.)				
Linear-discounting model ($\theta = 0.86$, $\beta = 0.81$)	.084	.715	.111	.264
Constant-w model ($\theta = 1.34$, $w = 0.54$)	.085	.708	.124	.104
Noncompensatory model ($m = 8.67$, $b = 4.99$)	.093	.648	.150	.256
Experiment 4 (freq.)				
Linear-discounting model ($\theta = 1.24$, $\beta = 0.79$)	.095	.744	.086	.232
Constant-w model ($\theta = 1.64$, $w = 0.64$)	.095	.743	.099	.001
Noncompensatory model ($m = 10.28$, $b = -7.07$)	.112	.642	.171	.237

Note. Models are tested in their ability to account for the full set of 96 mean pattern judgments per experiment, as well as for the corresponding set of 32 mean total probability or frequency assigned to the three flu strains given a particular symptom pattern, using the mean similarity ratings from Experiment 5 as a measure of evidential support. RMSE = root-mean-square error; m = slope; b = y-intercept.

Results of fitting the linear-discounting model to the probability and frequency judgments are presented in Table 9. Although closer fits have been achieved using the model in some previous studies (see Koehler et al., 1997), the model did account for more than 70% of the variance in the set of observed judgments in each dataset.

Once the model was fit to the probability and frequency judgment data, its predictions regarding the total probability T assigned to the three flu strains for a given symptom pattern were compared to the observed mean totals from Experiments 3 and 4. Fit statistics are presented in Table 9. The results are displayed graphically in Figures 1 and 2, in which the symptom patterns have been ordered by the number of present (positive) symptoms included in the pattern. The linear-discounting model appears to have at least a moderate ability to predict variability in T as a

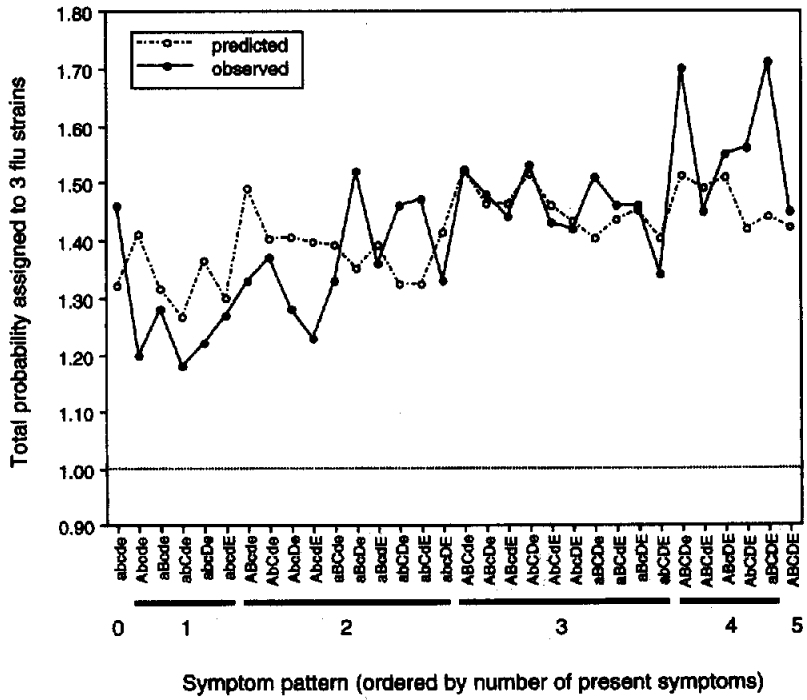

Figure 1. Total probability T assigned to the three flu strains in Experiment 3, by symptom pattern. Patterns are ordered by the number of positive (i.e., present) symptoms they include. Total probability predicted by the linear-discounting model, using support estimates from Experiment 5, is also shown.

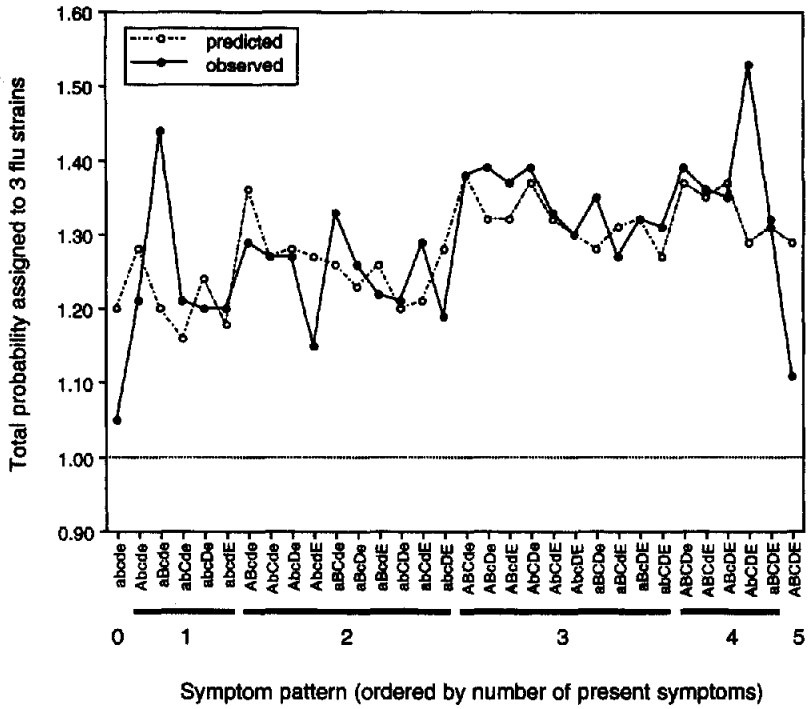

Figure 2. Total of derived probability T assigned to the three flu strains in Experiment 4, by symptom pattern. Patterns are ordered by the number of positive (i.e., present) symptoms they include. Total probability predicted by the linear-discounting model, using support estimates from Experiment 5, is also shown.

function of the symptom pattern displayed by the patient, though there is clearly still much room for improvement. It is worth noting in this regard that the model's predictions are based on an indirect fitting to the pattern judgment data rather than on a direct fitting to the set of T values; obviously, a better fit could be achieved if the set of T values were to be fitted directly. The focus here, however, is on the ability of a model of the individual pattern judgments to also predict the value of T.

The fit of the linear-discounting model can be compared to a *constant-w model* in which the degree of discounting of hypotheses included in the implicit residual hypothesis is unrelated to the support for the focal hypothesis (Koehler et al., 1997). This model provides a baseline against which to assess the usefulness of the linear-discounting model, by assessing how well a model can perform that assumes no systematic variation in the degree of discounting. The constant-w model also has two free parameters, θ and w, where w is the value of the constant discounting factor. As indicated in Table 9, the constant-w model achieved a fit to the judgment data that was only slightly worse than that achieved by the linear-discounting model. Large differences in performance between the two models are not to be expected given the number of features they share: Both models take into account the relative support for the focal and alternative hypotheses, and both assume that the alternatives lose support by being packed together in an implicit residual hypothesis.

The constant-w model, however, exhibited a much poorer ability to predict the total probability T for a given symptom pattern than did the linear-discounting model, as can be seen in Table 9. Despite the relatively small number of observations being fit, the difference in fit to the T values between the two models is marginally statistically significant for the probability judgment data of Experiment 3, $t(29) = 1.26$, $p < .11$, and is statistically significant for the frequency judgment data of Experiment 4, $t(29) = 3.04$, $p < .01$.

Finally, a *noncompensatory model* was also fit to the judgment data from Experiments 3 and 4. In this model, it is assumed that the judged probability (or frequency) of a particular hypothesis depends only on the support for that hypothesis, regardless of the extent to which the evidence also supports the alternative hypotheses. Support theory, by contrast, assumes that judged probability generally increases with the support for the focal hypothesis and decreases with the support for the alternative hypothesis, even if elementary hypotheses implicitly included in the residual acting as the alternative hypothesis lose support by being packed together. This assumption of compensatory judgments is incorporated in both the linear-discounting model and the constant-w model; the noncompensatory model offers a test of its validity.

The noncompensatory model was instantiated as a simple linear regression of the pattern judgments (from Experiments 3 and 4) on the associated support value for the focal hypothesis as measured by the corresponding similarity judgment from Experiment 5. Thus, like the linear-discounting model and the constant-w model, the noncom-

pensatory model has two free parameters, in this case m and b, representing the slope and intercept of the regression line, respectively. Table 9, which presents the fit of this model to the probability and frequency judgments, clearly indicates the relative inferiority of the noncompensatory model's fit to the data. It does not fit the pattern judgments as well as either of the other two models, and while in the correlational sense it appears to provide a comparable fit to the T values as that provided by the linear-discounting model, in absolute terms (measured by RMSE) its predictions deviate substantially from the observed values.

Assessment of the noncompensatory model's performance is relevant to a potential criticism of the use of similarity ratings from Experiment 5 as a measure of evidential support. Indeed, the choice of an appropriate task for direct rating of support can be complicated, with any particular task subject to some potential drawbacks (for discussion of this issue, see Koehler et al., 1997, p. 296). In this case, specifically, it could be argued that—despite experimental instructions—participants asked to make similarity judgments interpreted the task as one of judging the probability of the designated flu strain given the symptom pattern. On this interpretation, it would hardly be surprising that the resulting "support" measure consequently allows accurate fitting of the probability (and frequency) judgment data from Experiments 3 and 4 because, the argument goes, participants in the similarity-rating task are also making probability judgments. The relatively poor fit of the noncompensatory model to the judgment data, however, is inconsistent with this interpretation. If participants in Experiment 5 gave probability judgments instead of similarity ratings as instructed, then the simple scale transformation reflected in the noncompensatory model ought to provide a superior fit to the judgment data. The observation that the linear-discounting model and the constant-w model outperform the noncompensatory model suggests that the judgment data from Experiments 3 and 4 reflect a sensitivity to the support for the alternatives to the focal hypothesis not present in the similarity ratings. This is precisely the pattern of results that would be expected if participants in Experiment 5 did in fact interpret their task as one of judging similarity (a noncompensatory judgment task) rather than probability (a compensatory judgment task).

Summary. Collection of similarity judgments serving as a measure of evidential support allowed model fitting of the judgment data from Experiments 3 and 4, with informative results. The finding that the noncompensatory model was generally outperformed by the linear-discounting model and the constant-w model indicates that people's probability judgments are sensitive not only to the support for the focal hypothesis, but also to the support for the alternative hypothesis, as predicted by support theory. The finding that the linear-discounting model outperformed the constant-w model in fitting the total probability T assigned to the three flu strains indicates that variability across the symptom patterns in the extent to which alternatives are discounted by their inclusion in an implicit residual hypothesis is systematic. The form of the linear-discounting model—which assumes that the greater the support for the focal hypothesis, the greater is the extent to which support for alternatives is discounted in the residual—provides a more complete picture of the influence of cue conflict (and cue redundancy) observed in the previous experiments: Symptom patterns associated with high conflict (and redundancy) are perceived to provide greater evidential support for the hypotheses under consideration, which in turn invokes greater discounting.

General Discussion

There are two major findings from the present experiments. First, the total probability T assigned to the three categories consistently exceeded one in the context of a classification-learning task. Second, the value of T was influenced by cue conflict and redundancy but not by cue frequency, and was predictable from support theory using the linear-discounting model. The former finding has implications for the issue of when and why biases in probability judgment are likely to arise. The latter has implications for the study of how evidence is evaluated in the process of assessing support. These two issues are addressed in turn.

Interpreting Biases in Probability Judgment

When component hypotheses are implicitly packed together in the residual, they tend to lose support. As a result, the focal hypothesis enjoys a systematic advantage. This bias, observed in many previous studies in which judgments were based on general knowledge (for a review, see Tversky & Koehler, 1994), held across four classification-learning experiments in which the relevant knowledge is acquired in the laboratory. As is elaborated below, these results are inconsistent with arguments that systematic bias in human judgment is (a) attributable to biased selection of nonrepresentative items, and (b) eliminated by a frequentistic formulation.

Recently, debate has emerged over the interpretation of the overconfidence people display in many tasks (for reviews, see Harvey, 1997; Keren, 1991; Lichtenstein, Fischhoff, & Phillips, 1982; McClelland & Bolger, 1994; Wallsten & Budescu, 1983). A number of researchers (Björkman, 1994; Gigerenzer et al., 1991; Juslin, 1994) have offered ecological models of subjective probability calibration in the tradition of Brunswik (1943, 1955). These models focus on subjective cue validities that are acquired through experience in a given environment or reference class. They adopt a null hypothesis in which it is assumed that subjective cue validities are unbiased estimates of the actual validity of the cue in the environment. Any observed biases in laboratory studies, from this view, must be attributable to the manner in which test items are selected by the experimenter. In particular, if the experimenter selects a set of items for which the relevant cues have lower predictive validity than in the natural environment, the result will be an appearance of systematic bias (i.e., overconfidence). From this view, apparent systematic bias will be eliminated by representative sampling of test items from the natural reference class.

The classification-learning paradigm would seem to provide an ideal way in which to test such claims, as it allows complete experimental control over the learning environment. The present studies demonstrate that probability judgments continue to exhibit considerable subadditivity, even in an experimental design that avoids the problem of nonrepresentative sampling of test items. While the present experiments do not directly address the issue of subjective probability calibration, a recent study designed to investigate calibration did find substantial overconfidence using a classification-learning task (Yates, Lee, Shinotsuka, Patalano, & Sieck, 1998; see also McKenzie, 1997). Research using other methodologies (Griffin & Tversky, 1992; Juslin, Olsson, & Björkman, 1997) also corroborates the conclusion that representative sampling is not sufficient to eliminate overconfidence.

A second criticism of studies documenting judgmental biases comes from researchers arguing that the mind is designed for processing frequencies rather than probabilities (Brase et al., 1998; Cosmides & Tooby, 1996; Gigerenzer et al., 1991; Gigerenzer & Hoffrage, 1995). This view was elaborated in some detail in the introduction to Experiment 4, which provides a strong test of claims regarding the existence of a modular cognitive algorithm designed for optimal processing of frequencies. The design of this experiment ought to have been ideal for the operation of such an algorithm, as the evidence serving as input was presented in the form of a sequence of single cases, and the judgments required as output were elicited in the form of absolute frequency estimates. Nonetheless, the resulting judgments exhibited systematic subadditivity, demonstrating that elicitation of absolute frequencies is not sufficient to eliminate this particular form of judgmental bias (see also Rottenstreich & Tversky, 1997).

The frequentistic formulation used in Experiment 4 did reduce the degree of subadditivity observed in people's judgments relative to that found in Experiment 3 using a probabilistic formulation. This result is consistent with the general claim of Tversky and Kahneman (1983) that a frequentistic formulation can yield more extensional judgments. Tversky and Koehler (1994) suggested, specifically, that a frequentistic formulation reduces subadditivity by encouraging the judge to unpack implicit disjunctions to a greater extent than is typically done in judgments of probability. Asking the judge in Experiment 4 to assess the number of patients with a given pattern of symptoms who have a designated flu strain, out of the total number of patients with that pattern of symptoms, may have encouraged greater consideration of patients with that symptom pattern suffering from specific flu strains other than the designated flu strain. The frequentistic formulation does not generally produce complete unpacking of the residual hypothesis. As a result, frequency judgments are still subadditive, but less so than probability judgments.

Assessing Evidential Support

On the basis of the results of the experiments reported in this article and in related work, it is possible to offer a preliminary characterization of some apparently central principles governing the assessment of evidential support, using support theory as a guiding theoretical framework. A common theme of these principles, elaborated below, is a tendency toward reducing the complexity of assessing the implications of a multicue body of evidence for a set of hypotheses. To make these proposed principles more concrete, they will be illustrated with a running example. Consider the task of judging the probability of Flu Strain 1, $P(F_1, \overline{F}_1)$, given the symptom pattern AbCDe. According to support theory, this requires an assessment of the support for Flu Strain 1 and for its complement provided by the symptom pattern.

The first principle concerns the representation of hypotheses.

Principle 1: Composite residual formation. When a focal elementary hypothesis (e.g., F_1) is pitted against all of its alternatives taken together (e.g., F_2 and F_3), the alternatives are packed together to form an implicit disjunction referred to as the *residual hypothesis.* Koehler et al. (1997) suggest that in assessing the support for the alternatives to the focal hypothesis, people first form a composite representation of the alternatives in the form of an implicit residual hypothesis (e.g., "not F_1"), and then evaluate the support provided by the evidence for the resulting composite hypothesis.

An alternative process would be to assess the evidential support for each component hypothesis included in the residual, and then aggregate the support across all such component hypotheses to arrive at an overall assessment of the support for the residual hypothesis. Given a large set of elementary hypotheses and a complex body of evidence, this alternative process would quickly become computationally difficult, and it retains information that may very well be irrelevant to the judge regarding the distribution of support among specific alternatives to the focal hypothesis. Formation of a composite residual hypothesis that is evaluated as a single entity can simplify the task of support assessment.

The cost of this process appears to be a systematic loss of support for hypotheses included implicitly in the residual, relative to what they would have received if assessed individually. Even if there are compelling pieces of evidence supporting its components, there may be no single compelling piece of evidence supporting the composite residual hypothesis as a whole (Brenner & Koehler, 1999). In this manner, the absence of a single piece of evidence supporting the residual may put it at a substantial disadvantage in terms of support assessment, in much the same way that a set of disjunctive reasons for a choice option tends to have less impact on the option's attractiveness than does a single coherent reason (Tversky & Shafir, 1992).

The second and third principles concern the representation of evidence.

Principle 2: Evidence decomposition. When possible, a complex body of evidence is decomposed and evaluated in a piece-by-piece manner. That is, rather than assessing the implications of the configuration or pattern of symptoms taken as a whole (e.g., pattern AbCDe), each cue's (e.g., A's) contribution to the support for a hypothesis is assessed

individually. The focus on individual cues may represent a kind of initial default strategy that could be supplemented with analysis of configural cue information if such information proved to be sufficiently valuable (cf. Castellan & Edgell, 1973; Edgell, 1978, 1980; Edgell & Castellan, 1973; Edgell & Roe, 1995).

Of course, not all bodies of evidence will necessarily lend themselves to the kind of simple decomposition that is likely to have occurred in the present experiments. Sometimes, discrete features may be combined and perceived as a single piece of evidence (e.g., Edgell & Morrissey, 1992). For example, the presence of the symptom *cough* along with the absence of the symptom *chest congestion* may be combined into the single piece of evidence *dry cough.* Thus, the suggestion that a body of evidence may be evaluated on a piece-by-piece basis needs to take into account "pieces" or "chunks" of evidence as represented by the judge.

Evidence decomposition, too, can reduce the complexity of the support assessment process (cf. Hogarth & Einhorn, 1992). If, for example, participants kept track of the relationship between each possible symptom pattern (of which there are 32 given five binary symptoms) and each of the three flu strains, they would have to monitor a total of 96 relationships. In contrast, by taking the symptoms one at a time, many fewer relationships need to be considered (at most, 30, though possibly fewer given the next principle). Another advantage of evidence decomposition in this respect is that updating of belief as new pieces of information are encountered is relatively straightforward: The implication of the new piece of evidence can be assessed on its own, without an entire reassessment of the modified body of evidence as a whole.

Principle 3: Cue presence/absence asymmetry. Given binary cues representing the presence or absence of a feature, assessment of support for a hypothesis appears to be based primarily on present cues. Given symptom pattern AbCDe, for example, it is assumed that the support for a particular hypothesis is assessed primarily on the perceived implications of present Symptoms A, C, and D. This appears to be the case even when, as in the present experiments, cue absence is explicitly denoted by the presence of a label (e.g., *no cough*). As described in the introduction, this assumption is consistent with previous findings in studies of classification learning (e.g., Estes et al., 1989) and in the broader study of judgment under uncertainty (e.g., Kao & Wasserman, 1993). A focus on present cues is also a key component of one of Klayman and Ha's (1987) positive test strategies, called the *positive target test,* in which instances or cases known to have a target property are examined to see whether they fit a hypothesized classification rule.

The focus on present cues rather than absent cues seems intuitively reasonable, as any body of evidence can be characterized in terms of a finite set of present features but a potentially infinite set of absent features. For example, even assuming that only a limited (but presumably large) set of symptoms is considered generally relevant for medical diagnosis, the number of symptoms not exhibited by a patient will greatly exceed the number of symptoms the patient does exhibit.

Taken together, Principles 2 and 3 provide a substantial reduction in the computational requirements of the support assessment process. Recall that if symptom patterns were to be assessed in their entirety, in a configural manner, a total of 96 relationships would have to be monitored in the 5-symptom, 3-flu-strain case. By comparison, if each symptom pattern is evaluated in a symptom-by-symptom manner, and if only the implications of cue presence are assessed, the number of monitored relationships is reduced to 15. If the judge observes large, systematic correlations among features, or significant violations of conditional independence, of course, he or she might be more inclined to consider at least some subsets of cues together, which would necessarily increase the number of relationships that would have to be monitored.

The next three principles concern general properties of the support assessment process itself, which is a joint function of hypothesis and evidence representation.

Principle 4: Noncompensatory support assignment. The support for a hypothesis reflects only those aspects of the evidence which directly implicate that hypothesis. Evidence that indirectly relates to the likelihood of the hypothesis, by implicating one or more of its alternatives, is not generally reflected in the support for the indirectly affected hypothesis. Given symptom pattern AbCDe, for example, suppose that the judge views the presence of Symptom A as directly implicating F_1 but sees the presence of Symptom C as only indirectly relevant in the sense that it directly implicates a competing hypothesis, F_3. It is assumed that the presence of Symptom A will contribute (positively) to the support for F_1. By contrast, the presence of Symptom C is assumed not to influence the support for F_1. The support assessment process is thus said to be noncompensatory, meaning that it is insensitive to the degree to which the evidence supports alternative hypotheses.

Support theory provides a mechanism by which probability judgments can be compensatory even if the underlying support assessment process is not, namely the representation (1) of probability as normalized support for the focal and alternative hypotheses. Evidence that implicates an alternative hypothesis may not influence the support for the focal hypothesis, but will nonetheless reduce its judged probability by increasing the support for its alternative. (Hence the failure of the noncompensatory probability model to account for the probability and frequency judgments from Experiments 3 and 4.)

When a focal elementary hypothesis is pitted against all of its alternatives taken together as a residual hypothesis, however, the subadditive relationship between the support for the residual and that for its components when assessed individually results in partially noncompensatory probability judgments. When a piece of evidence directly implicates the focal hypothesis, the support for the focal hypothesis is adjusted accordingly, but when the implicated hypothesis is included implicitly in the residual hypothesis, its influence is discounted. Thus, the noncompensatory nature of the support assessment process "leaks" into the resulting probability judgments, as illustrated by the enhancement effect (Brenner & Koehler, 1999; Koehler et al., 1997; Tversky &

Koehler, 1994). A number of previous researchers (e.g., Robinson & Hastie, 1985; Teigen, 1983; Van Wallendael, 1989; Van Wallendael & Hastie, 1990) have also reported that likelihood judgments made under conditions of uncertainty appear to be systematically noncompensatory. As some of these researchers have noted, a noncompensatory process has the advantage of computational simplicity. In assessing the evidential support for a particular hypothesis, only those aspects of the evidence that directly implicate the hypothesis need to be considered. Furthermore, the same support value $s(F_1)$ can be used when judging $P(F_1, F_2)$, $P(F_1, F_3)$, and so on, assuming all the hypotheses in question correspond to events in a fixed sample space. (See Koehler, 1996, for elaboration and extension of this observation.)

Principle 5: Diagnosticity-based support assignment. Results of the present experiments suggest that the support assessment process is based in large part on the perceived diagnosticity of the individual cues comprising the body of evidence. In these studies at least, the single most important determinant of a present cue's contribution to the perceived support for a hypothesis is the extent to which the cue's presence is perceived as differentially associated with that hypothesis. Mere association of the cue with the hypothesis, due to the cue's overall prevalence in the environment, is not perceived as providing strong support for the hypothesis, as indicated by the consistent null effect of cue frequency in the present experiments. The sensitivity of support assessments to evidential diagnosticity will help to ensure at least some degree of judgmental accuracy despite the many simplifications in evidential reasoning reflected in Principles 1–4.

How can this claim that support assessments are sensitive to evidential diagnosticity be reconciled with the earlier suggestion that support assessments are noncompensatory? The diagnosticity of a piece of evidence, after all, depends on the extent to which that piece of evidence is associated with alternative hypotheses. Note, however, that the type of diagnosticity in question is that of a single piece of evidence, not that of the entire body of evidence upon which the probability judgment is based. For example, Body of Evidence ABC, when taken as a whole, is nondiagnostic (in Experiments 3–5) in that the (posterior) probabilities of the hypotheses under evaluation are unchanged from what they would be in the absence of any evidence at all (i.e., their prior probabilities). Taken individually, however, each piece A, B, and C has diagnostic value associating it with F_1, F_2, and F_3, respectively. Thus, on a piece-by-piece assessment of support (Principle 2), Body of Evidence ABC increases the support for all three hypotheses, producing enhanced subadditivity.

Principle 6: Support accumulation. In the present experimental context at least, the process of support assessment appears to be best characterized as one in which support is accumulated, over individual pieces of evidence in the evidence body, or over time as further evidence is encountered. Clearly, there are cases in which a piece of evidence can decrease the perceived support for a hypothesis; this happens, for example, when exculpatory evidence such as an alibi eliminates a suspect from a criminal investigation. Nonetheless, it is suggested, there may be a systematic tendency for evidence to increase rather than decrease the perceived support for a hypothesis.

The influence of cue redundancy in Experiments 3 and 5 (though notably absent in Experiment 4) is consistent with this claim: The presence of additional nondiagnostic pieces of evidence (i.e., Symptoms D and E) appeared to enhance support for the designated hypothesis. This observation suggests that—holding fixed the diagnostic implications of the evidence—a hypothesis tends to receive more support from a body of evidence when it includes more present features or cues, as would be expected if each piece of evidence has a tendency to add to the support for a hypothesis.

Such an influence is also apparent in comparing the results of Experiments 1 and 3, which used four and five binary symptoms, respectively. On average, then, symptom patterns in Experiment 3 have a greater number of present symptoms than symptom patterns in Experiment 1. (Of course, they have a greater number of absent symptoms on average as well.) As expected if the number of present cues has an influence, the mean value of T in Experiment 3 ($M = 142\%$) is substantially greater than that found in Experiment 1 ($M = 120\%$).

Additional empirical findings consistent with a support-accumulation process are reported by Robinson and Hastie (1985, Experiment 3), who presented participants with murder mysteries, one clue at a time, and elicited probability judgments of guilt for each suspect following each clue. Participants' judgments were generally noncompensatory, exhibiting a pronounced tendency to adjust the probability of the implicated suspect only, consistent with Principle 4. More critically in the context of Principle 6, the judged probability of guilt increased in a linear fashion with the number of clues indicating the suspect's guilt, but did not exhibit comparable decreases as the number of clues indicating innocence increased. As would be expected assuming a support accumulation process, clues appeared to be more likely to increase than to decrease support for the hypothesis that a given suspect was guilty.

Postulation of Principles 1–6 goes well beyond the data reported in this article and that available from previous research, and will clearly require further investigation. Furthermore, these principles are intended to capture some apparent general tendencies in probabilistic reasoning; they are bound to be subject to some exceptions. Identification of conditions under which the principles outlined above hold, or fail to hold, should be highly informative. A related issue concerns the environmental conditions under which adherence to these principles yields reasonable accuracy or substantial error (cf. Klayman & Ha, 1987; McKenzie, 1994).

Indeed, there are already a few notable caveats regarding the principles above that can be found in previous research. McKenzie (1998, 1999), for example, has shown that the degree to which probability judgments assigned to complementary hypotheses are compensatory or noncompensatory depends systematically on the manner in which the relationship between the hypotheses and cues is learned. As another

example, Hogarth and Einhorn (1992) considered task characteristics that might contribute to the extent to which evidence is assessed holistically versus one piece at a time. In some cases the judge may assess the implications of one piece of evidence in light of another piece from the same body of evidence, focusing on the "configural" relations among the set of cues (e.g., Edgell & Roe, 1995) and possibly evaluating evidence by reference to a set of stored exemplars (e.g., Estes, 1986; Medin & Schaffer, 1978; Nosofsky, 1986).

Models of classification learning are intended to address precisely these kinds of issues. As currently formulated, however, many classification-learning models are unable to account for the two main findings of the present study: Probability judgments are systematically subadditive, such that the total probability T assigned to a set of three or more possibilities consistently exceeds one, and variance in T is predictable from evidential characteristics such as cue conflict. While most existing models of classification learning were originally developed to fit choice probabilities, which are necessarily additive, recent attempts to generalize such models to judged probability appear to have implicitly retained the additivity assumption (e.g., Estes et al., 1989; Gluck & Bower, 1988; Nosofsky et al., 1992). Such an assumption is inconsistent with the results presented in this article.

Systematic violations of additivity may have gone unnoticed in the classification learning literature in part because the typical experiment in this area involves learning to discriminate between a pair of categories. Under these conditions, in which there are only two complementary hypotheses, support theory predicts additive probability judgments (because no residual hypothesis needs to be formed), a prediction consistent with observations from a number of researchers that probability judgments appear more compensatory in the two-hypothesis case (e.g., Robinson & Hastie, 1985; Teigen, 1983; Van Wallendael & Hastie, 1990). It is only when there are three or more mutually exclusive hypotheses that support theory predicts $T > 1$. Generalization of classification-learning models to learning involving more than two categories may pose greater difficulties than has been previously assumed.

References

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

Brase, G. L., Cosmides, L., & Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole-object representations in judgment under uncertainty. *Journal of Experimental Psychology: General, 127,* 3–21.

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.

Brenner, L. A., & Rottenstreich, Y. (1999). Focus, repacking, and the judgment of disjunctive hypotheses. *Journal of Behavioral Decision Making, 12,* 141–148.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3.

Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review, 50,* 255–272.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62,* 193–217.

Castellan, N. J., Jr. (1977). Decision making with multiple probabilistic cues. In N. J. Castellan, Jr., D. B. Pisoni, & G. R. Potts (Eds.), *Cognitive theory* (Vol. 2, pp. 117–147). Hillsdale, NJ: Erlbaum.

Castellan, N. J., Jr., & Edgell, S. E. (1973). An hypothesis generation model for judgment in nonmetric multiple-cue probability learning. *Journal of Mathematical Psychology, 10,* 204–222.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58,* 1–73.

Dube-Rioux, L., & Russo, J. E. (1988). An availability bias in professional judgment. *Journal of Behavioral Decision Making, 1,* 223–237.

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.

Edgell, S. E., & Castellan, N. J., Jr. (1973). Configural effect in multiple-cue probability learning. *Journal of Experimental Psychology, 100,* 310–314.

Edgell, S. E., & Morrissey, J. M. (1992). Separable and unitary stimuli in nonmetric multiple-cue probability learning. *Organizational Behavior and Human Decision Processes, 51,* 118–132.

Edgell, S. E., & Roe, R. M. (1995). Dimensional information facilitates the utilization of configural information: A test of the Castellan–Edgell and the Gluck–Bower models. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1495–1508.

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.

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole. *Journal of Personality and Social Psychology, 66,* 633–645.

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.

Flavell, J. H. (1963). *The developmental psychology of Jean Piaget.* New York: Van Nostrand.

Fox, C. R., Rogers, B., & Tversky, A. (1996). Options traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty, 13,* 5–19.

Fox, C. R., & Tversky, A. (1998). A belief-based account of decision under uncertainty. *Management Science, 44,* 879–895.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102,* 684–704.

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

Gluck, M. A., & Bower, G. H. (1988). From conditioning to

category learning: An adaptive network model. *Journal of Experimental Psychology: General, 117,* 227–247.

Griffin, D., & Buehler, R. (1999). Frequency, probability, and prediction: Easy solutions to cognitive illusions? *Cognitive Psychology, 38,* 48–78.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24,* 411–435.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences, 1,* 78–82.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1–55.

Jenkins, H. H., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. *Psychological Monographs, 79* (1, Whole No. 79).

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

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

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. *TIMS Studies in Management Science, 12,* 313–327.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11,* 143–157.

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.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217–273.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94,* 211–228.

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 375–388). New York: Springer-Verlag.

Koehler, D. J. (1996). A strength model of probability judgments for tournaments. *Organizational Behavior and Human Decision Processes, 66,* 16–21.

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

Lewandowsky, S. (1995). Baserate neglect in ALCOVE: A critical reevaluation. *Psychological Review, 102,* 185–191.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: Cambridge University Press.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgments. *Psychological Review, 106,* 210–214.

McClelland, A. G. R., & Bolger, F. (1994). The calibration of subjective probabilities: Theories and models 1980–94. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 453–482). Chichester, England: Wiley.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology, 26,* 209–239.

McKenzie, C. R. M. (1997). Underweighting alternatives and overconfidence. *Organizational Behavior and Human Decision Processes, 71,* 141–160.

McKenzie, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 771–792.

McKenzie, C. R. M. (1999). (Non)complementary updating of belief in two hypotheses. *Memory & Cognition, 27,* 152–165.

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

Mehle, T., Gettys, C. F., Manning, C., Baca, S., & Fisher, S. (1981). The availability explanation of excessive plausibility assessments. *Acta Psychologica, 49,* 127–140.

Newman, J. P., Wolff, W. T., & Hearst, E. (1980). The feature-positive effect in adult human subjects. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 630–650.

Norton, G. R., Muldrew, D., & Strub, H. (1971). Feature-positive effect in children. *Psychonomic Science, 23,* 317–318.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

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.

Pelham, B. W., Sumarta, T. T., & Myaskovsky, L. (1994). The easy path from many to much: The numerosity heuristic. *Cognitive Psychology, 26,* 103–133.

Peterson, D. K., & Pitz, G. F. (1988). Confidence, uncertainty, and the use of information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 85–92.

Piaget, J., & Inhelder, B. (1941). *Le développement des quantités chez l'enfant* [The development of quantities in the child]. Neuchâtel, Switzerland: Delachaux et Niestlé.

Redelmeier, D., Koehler, D. J., Liberman, V., & Tversky, A. (1995). Probability judgment in medicine: Discounting unspecified alternatives. *Medical Decision Making, 15,* 227–230.

Reeves, T., & Lockhart, R. S. (1993). Distributional vs. singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General, 122,* 207–226.

Robinson, L. B., & Hastie, R. (1985). Revision of beliefs when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 443–456.

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.

Shafer, G. (1976). *A mathematical theory of evidence.* Princeton, NJ: Princeton University Press.

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.

Shanks, D. R. (1990). Connectionism and the learning of probabilistic concepts. *Quarterly Journal of Experimental Psychology, 42A,* 209–237.

Shanks, D. R. (1991). Categorization by a connectionist network. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 433–443.

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

Teigen, K. H. (1974a). Overestimation of subjective probabilities. *Scandinavian Journal of Psychology, 15,* 56–62.

Teigen, K. H. (1974b). Subjective sampling distributions and the additivity of estimates. *Scandinavian Journal of Psychology, 15,* 50–55.

Teigen, K. H. (1983). Studies in subjective probability: III. The unimportance of alternatives. *Scandinavian Journal of Psychology, 24,* 97–105.

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.

Tversky, A., & Shafir, E. (1992). The disjunction effect in choice under uncertainty. *Psychological Science, 3,* 305–309.

van der Pligt, J., Eiser, J. R., & Spears, R. (1987). Comparative judgments and preferences: The influence of the number of response alternatives. *British Journal of Social Psychology, 26,* 269–280.

Van Wallendael, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior and Human Decision Processes, 43,* 385–405.

Van Wallendael, L. R., & Hastie, R. (1990). Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing. *Memory & Cognition, 18,* 240–250.

Wallsten, T. S., & Budescu, D. V. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science, 29,* 151–173.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74,* 89–117.

Received June 19, 1995
Revision received June 23, 1999
Accepted June 28, 1999 ■