Probability Judgment in Three-category Classification Learning

Derek J. Koehler (dkoehler@watarts.uwaterloo.ca)
Department of Psychology, University of Waterloo
Waterloo, Ontario N2L 3G1 CANADA

Abstract

People tend to give subadditive probability judgments when asked to assess each in a set of three or more exclusive hypotheses. The degree of subadditivity in such judgments is determined in large part by the evidence upon which the judgments are based, but the characteristics of the evidence that influence subadditivity have yet to be fully specified. In the present experiments, this issue was addressed using a classification learning task, in which the relationship between the evidence and the hypotheses under consideration can be controlled experimentally. Two potential evidential influences on subadditivity--cue conflict and cue frequency--are distinguished and tested in three experiments. The results indicate that (a) people's probability judgments are systematically subadditive--in violation of standard probability theory--even when the judgments are based on cues learned within the experimental context, contrary to the predictions of "ecological" theories of human judgment which attribute such biases to nonrepresentative item selection; and (b) cue conflict has a reliable influence on the degree of subadditivity exhibited in probability judgments.

There is substantial evidence that people's probability judgments are nonextensional, that is, not consistent with the rules of set inclusion. Recently, a descriptive theory of probability judgment called support theory (Tversky & Koehler, 1994) has been developed to account for these findings. Support theory makes two basic assumptions. The first is that judged probability reflects the relative support for the focal and alternative hypotheses:

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

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. 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 and B refer to descriptions of events, called *hypotheses*, rather than to the events themselves, as in standard probability theory.

Support theory distinguishes between explicit disjunctions, which list their components, and implicit disjunctions, which do not. Support theory's second assumption is that if H is an implicit disjunction (e.g., homicide) that refers to the same event as an explicit disjunction of exclusive hypotheses H_a and H_s (e.g., homicide by an acquaintance or homicide by a stranger, denoted $H_a \vee H_s$), then

$$(2) \qquad s(H) \leq s(H_a \vee H_s) \leq s(H_a) + s(H_s).$$

That is, the support of the implicit disjunction H is less than or equal to that of the explicit disjunction $H_a \vee H_s$, 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 hypothesis into its components can only increase its support, and hence its judged probability (cf. Fischhoff, Lichtenstein, & Slovic, 1978). The relationship between the support of the implicit disjunction and that of its components is said to be subadditive, in the sense that the whole receives less than the sum of its parts.

Support theory implies that, whenever a single well-specified hypothesis is evaluated relative to all of its alternatives taken as a group (referred to as a "catchall" or *residual* category), the specified hypothesis will be given greater weight than if it had been included implicitly in the residual category. Consider an example with three hypotheses: A, B, and C. When a person is asked to judge the probability of hypothesis A, according to support theory, the probability judgment is determined by the evidential support for hypothesis A normalized relative to that for its complement. In this case its complement is an implicit disjunction of hypotheses B and C. Support theory assumes that the implicit representation of the alternative hypotheses decreases their support relative to that of A, thereby increasing A's judged probability. If separate judgments are obtained of the probability of hypotheses A, B, and C, the total probability assigned to the three is predicted to exceed one, in violation of standard probability theory. The degree of subadditivity in this case can be measured by the extent to which the total exceeds one.

The degree of subadditivity observed depends on a number of factors (see Tversky & Koehler, 1994), including the compatibility of the evidence with each of the hypotheses under consideration. For example, in one experiment (Koehler, Brenner, & Tversky, 1997, Exp. 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 (measured by the total probability 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*.

While the notion of "compatibility" between evidence and hypotheses serves to summarize a number of manipulations observed to influence subadditivity, the exact characteristics of the evidence controlling subadditivity 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. This was accomplished in the present investigation through the use of a simulated medical diagnosis task, which has been used in much of the recent work on classification learning (e.g., Estes, Campbell, Hatsopoulos, & Hurwitz, 1989; Gluck & Bower, 1988; Nosofsky, Kruschke, & McKinley, 1992). In this task, participants are presented with a set of symptoms ("cues" which serve as evidence) reported by a "patient" and are asked to guess which of a set of possible diseases (typically two) the patient might have. Participants are presented with a large number of patients; after each guess 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.

This task was used to investigate two possible interpretations of evidential compatibility underlying the enhancement effect. The first possibility involves overall cue frequency: Subadditivity may increase with the frequency of presentation during learning of the cue used as the basis of judgment. That is, if some cues simply occur more often in conjunction with all of the categories than do others, presentation of these cues for judgment may yield greater subadditivity than less frequently presented cues. The second sense in which enhancement may operate involves the degree of conflict among a set of cues. Research on enhancement (Koehler et al., 1997; Tversky & Koehler, 1994) suggests that subadditivity may be increased by the introduction of evidence that has mixed or conflicting implications (e.g., Peterson & Pitz, 1988). In the current context this possibility can be examined by analyzing different patterns of cues. Increased subadditivity would be expected for those patterns that imply or support more than one category or hypothesis.

In addition, the present set of experiments affords an opportunity to test two competing theories of human judgment. Support theory, with its origins in the heuristics and biases research programme of Tversky and Kahneman (e.g., 1974), assumes that the inferential mechanisms underlying probability judgment often produce reasonably accurate judgments but also cause systematic biases under certain circumstances. In contrast, some researchers (Björkman, 1994; Gigerenzer, Hoffrage, & Kleinbölting, 1991; Juslin, 1994) have recently suggested that judgmental biases observed in experimental settings arise because participants attempt to apply cues they have learned accurately from experience to a set of items selected by the experimenter that is non-representative of the environment in which the cue-outcome relationship was originally learned. This claim is tested in the present experiments. In a classification learning experiment, the environment in which the cue-outcome relationship holds is defined by and learned during the training sequence of the laboratory task itself. Thus the "ecological" approach leads to the prediction that the systematic subadditivity implied by support theory should not be observed in the present studies, in which nonrepresentative item selection is not an issue.

Experiment 1

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 as instructed, 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 upper-case letters, e.g., *COUGH*) or absent (in lower-case letters, e.g., *no cough*) on the medical chart. Each patient was to be classified as having one of three types of flu strains, simply labeled #1, #2, and #3.

Design. As in Estes et al. (1989) and Nosofsky et al. (1992), all participants were presented with an identical training sequence, consisting of 240 trials. This sequence was constructed by first randomly choosing one of the three flu strains (with equal probabilities), and then choosing the four symptoms (independently) with conditional probabilities yielding 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, each symptom, taken on its own, has the same diagnosticity. That is, given the symptom, 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 learning trials) and the other two flu strains decrease in probability to 20% each. Figure 1 indicates the mapping between the four symptoms and the flu strain with which each is most strongly associated.

The actual symptom label (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. Unlike the training sequence, which was the same for all participants, the order in which the subsequent 48 pattern judgments were made was determined randomly for each participant.

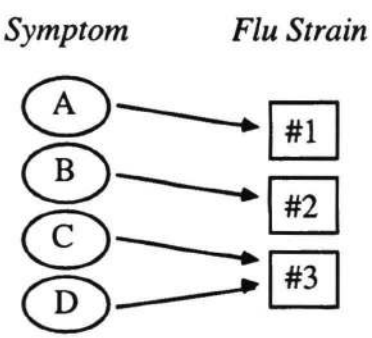

Figure 1: Schematic diagram of symptom-flu strain mapping in Experiments 1 and 2.

Procedure. Participants were told that they would be taking part in a simulated medical judgment task. They were told they would be presented with a series of 240 patients, each of whom was subsequently found (via a blood test) to have one of three influenza strains. 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 be guessing essentially at random, but as they saw more patients they should begin to have some sense of which symptoms go with which flu strains. They 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.

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. Participants were asked to make such judgments for all 48 possible combinations of the 16 different symptom patterns with the 3 different flu strains.

Results and Discussion

Learning Data. Over participants, average accuracy across the 240 training trials was 55%, a figure substantially greater than that expected by chance. All participants included in the sample achieved above-chance accuracy. To determine whether learning was at asymptote by the end of the 240 training trials, average percent correct 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. Participants' performance was no longer improving after the first 60 or so training trials, suggesting that by the end of the training phase participants had learned all they could about the category structure.

Pattern Judgment Data. Figure 2 displays the mean probability assigned to each flu strain for the 16 possible symptom patterns (present symptoms are denoted with uppercase letters, absent symptoms with lowercase). Participants' probability judgments were strongly related to the normative probabilities used to construct the training sequence. The correlation between the set of mean pattern judgments and the normative values is 0.93, showing that participants were able to translate what they had learned during the training sequence into reasonably accurate probability judgments. As predicted by support theory,

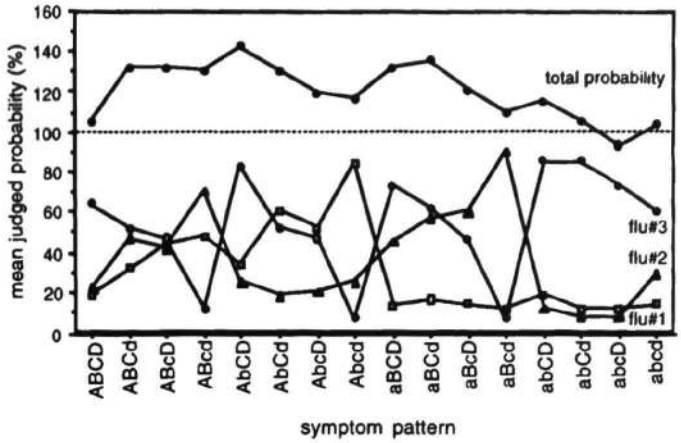

Figure 2: Mean judged probability of each flu strain and their total for each of the possible symptom patterns.

however, the probability judgments were clearly subadditive: The total probability assigned to the three possible flu strains consistently exceeded 100%, with an average total of 120% (see Figure 2).

The degree of subadditivity varied substantially over the various symptom patterns, allowing a test of the role of cue frequency and cue conflict on the degree of subadditivity. First consider cue conflict. As a simple comparison, the average total probability for those patterns with symptoms present that imply either zero or one flu strains ($n = 6$; e.g., abcd, Abcd, abCD) was compared with the average for those patterns implying two or more different flu strains ($n = 10$; e.g., ABcd, ABCd, ABCD). As predicted by the cue conflict interpretation of enhancement, the average total was significantly higher in the latter case ($M = 128\%$) than in the former ($M = 107\%$), $t\ (254) = 4.86$, $p < .001$.

For the pattern judgments, the cue frequency interpretation of enhancement can be tested in two ways. The simplest way is to consider only the patterns with a single present symptom (i.e., Abcd, aBcd, abCd, abcD). The mean total probabilities for these four patterns are 116%, 109%, 105%, and 93%, respectively, showing that subadditivity did increase with cue frequency. The contrast between the A and B symptoms and the C and D symptoms is statistically significant, $t\ (62) = 2.04$, $p < .05$. A more complicated analysis involves comparing the average totals for the eight patterns that include each symptom so that, for example, the pattern ABcd is counted as an A pattern and a B pattern but not as a C or D pattern. By this analysis the average totals for patterns including symptoms A through D are 126%, 124%, 124%, and 120%, respectively, again consistent with the cue frequency interpretation.

Experiment 2

The first experiment revealed substantial subadditivity in probability judgments elicited after learning, even though the task involved only three categories and gave a frequentistic interpretation to the response scale (cf. Gigerenzer et al., 1991). The degree of subadditivity was affected by both cue conflict and cue frequency. It could be argued, however, that had probability judgments been

elicited within the learning context, instead of after learning had taken place, the effects of cue conflict and frequency or even the general observation of subadditivity might have been eliminated. The feedback provided after each trial, for example, might draw participants' attention to the fact that their probability judgments are generally too high and hence eliminate 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.

Method

Participants. Participants were 34 prospective psychology undergraduate majors at University College London, who participated as part of a laboratory demonstration. Data from 3 of these participants were dropped as their learning performance was only marginally better than that expected by chance, leaving a total of 31 participants.

Design. Participants received the same training sequence as in Experiment 1, but assigned a probability to a designated flu strain rather than choosing which of the three flu strains they thought was most likely on each trial. 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 training trial. The flu strain 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. Because participants were giving what were referred to as pattern judgments in Experiment 1 on every trial of Experiment 2, participants were not asked to give final pattern judgments at the end of the learning 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). Probability judgments were made on a scale running from 0% to 100% in increments of 10%.

Results and Discussion

Learning Performance. In this experiment a more complicated analysis is necessary because participants judged the probability of a designated flu strain rather than choosing the flu strain they thought was most likely. A standard squared error measure was computed for each participant, which assumed a value of $(1 - p)^2$ on trials in which the designated event occurs and a value of p^2 on trials in which the designated event does not occur. Given chance performance (i.e., in the absence of any learning), the value of this error measure depends on the participant's response distribution. To adjust for this, a corrected performance score was computed for each participant by first calculating the expected value of chance performance given that participant's response distribution, and then subtracting the resulting value from the participant's actual performance score to obtain a measure of performance above that expected by chance. The resulting measure will be referred to as *corrected performance*.

All participants performed better than chance, that is, had positive corrected performance measures. The mean corrected performance value was 26.3 ($\underline{SD}$ = 11.8). All subsequent analysis is based on mean data averaged over participants within a given target group ($\underline{n}$ = 9, 11, and 11 for the three groups). As in the previous experiment, mean learning performance was examined for the four sequential sets of 60-trial blocks. The mean correct 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.

Pattern Judgments. 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. Figure 3 displays the mean judgment assigned to each flu strain, and their total, for each of the 16 possible symptom patterns. The correlation between the mean pattern judgments and the corresponding normative values was 0.92, which is essentially identical to that obtained in the first experiment. The correlation between the pattern judgments obtained in Experiments 1 and 2 is 0.95.

Once again, the probability judgments were substantially subadditive for all 16 symptom patterns. The (unweighted) mean total probability assigned to the three possible flu strains is 124%, which is slightly greater than the comparable value of 120% for the pattern judgments of

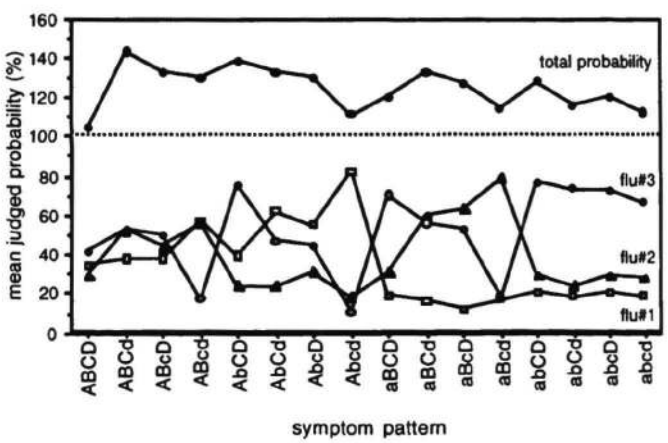

Figure 3: Mean judged probability of each flu strain and their total for each of the possible symptom patterns in Experiment 2.

the first experiment. Participants' judgments were consistently subadditive, then, even when feedback regarding the actual outcome was given after every judgment. Indeed, comparison with the pattern judgments of Experiment 1 suggests that the feedback did nothing at all to decrease the degree of subadditivity in the pattern judgments.

Cue conflict had a significant effect on the degree of subadditivity observed in the pattern judgments. Those patterns implying only a single flu strain (or none) received a weighted average total probability of 115% while those patterns implying two or three different flu strains received an average total of 132%, $t(178) = 6.03$, $p < .001$. Patterns including only a single present symptom (i.e., Abcd, aBcd, abCd, abcD) failed to show an effect of cue frequency, though it should be noted that the collection of judgments during learning allowed less time for effects of frequency to emerge.

Experiment 3

A final experiment assessed the effects of cue conflict and cue frequency using a different category structure than that used in the first two experiments. The new category structure was intended to completely separate testing of the two effects. In the resulting design, cue conflict could be tested using cues that were equated in terms of frequency, and cue frequency could be tested using cues that were completely nondiagnostic with regard to the outcome variable. As in the first experiment, participants made choice decisions rather than probability judgments on each of the training trials.

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.

Design and Procedure. Participants were presented with information regarding five symptoms, rather than four as in the previous experiments. Symptoms A, B, and C were equally diagnostic, and were associated with flu strains #1, #2, and #3, respectively. The likelihood 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. As in the first two experiments, then, the likelihood 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. Symptoms D and E were nondiagnostic and differed only in terms of their overall frequency. 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.

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. In contrast to the fixed training sequence used in the previous experiments, the order in which the 240 patients were presented in the training sequence was determined randomly for each participant to ensure the results were not attributable to some idiosyncrasy of the particular training sequence being employed. The introduction of a fifth symptom ("dizziness") increased the number of pattern judgments made by each participant 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% .

Results and Discussion

Learning Performance. Over participants, average accuracy across the 240 training trials was 47%. Participants were less accurate in the training phase of this experiment than they were in the previous two, as 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. 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.

Pattern Judgment Data. The correlation between the mean pattern judgments and the normative values over the 32 possible symptom patterns was 0.81. As predicted by support theory, the probability judgments were clearly subadditive: The total probability assigned to the three possible flu strains consistently exceeded 100%, with an average total of 142%. The degree of subadditivity observed for these judgments appears to be considerably greater than that of the previous experiments, perhaps because the inclusion of an additional symptom induced a greater sense of conflict or uncertainty.

The average total probability for those patterns with symptoms present that imply either zero or one flu strains ($n = 48$; e.g., Abcde, aBcDE, abcde) was compared with the average for those patterns implying two or more different flu strains ($n = 48$; e.g., ABcde, ABCde, ABCDe). As predicted by the cue conflict interpretation of enhancement, the mean total was significantly higher in the latter case ($M = 150\%$) than in the former ($M = 135\%$), $F(1, 15) = 13.30$, $p < .01$.

Recall that symptoms D and E were introduced to provide a test of cue frequency's role. Comparison of symptom patterns abcDe and abcdE revealed no significant difference, with mean total probabilities of 122% and 127%, respectively, $t(15) = 0.26$, ns. This difference is in the opposite direction of that predicted by the cue frequency interpretation of enhancement. Note, however, that this

analysis is based on only a single observation per participant. To overcome this problem, the symptom patterns were divided into four classes of 8 patterns each: de (neither D nor E); dE (E but not D); De (D but not E); and DE (both D and E). The average total probability assigned to these four classes was 134%, 139%, 146%, and 148%, respectively. These four classes differed significantly by an omnibus ANOVA, $F(3, 45) = 5.01$, $p < .01$; more importantly, the contrast between dE and De was at least marginally significant, $F(1, 45) = 2.85$, $p < .10$. There is some indication, then, that cue frequency plays a role even when the cues in question are completely nondiagnostic.

General Discussion

There are two major empirical findings in the present set of experiments. First, when more than two categories are used, people's probability judgments in the context of classification learning--as has been found in other domains--are substantially subadditive. Furthermore, use of the probability judgment task during (rather than after) the training sequence was insufficient to eliminate this effect: Participants continued to give subadditive judgments despite the provision of potentially corrective outcome feedback on every trial. Contrary to the claims of researchers such as Gigerenzer et al. (1991) and Juslin (1994), systematic biases in probability judgment are not necessarily eliminated by designs which exclude the possibility of non-representative item selection.

The second major empirical observation is that the degree of subadditivity in people's probability judgments varied substantially as a function of the evidence being used to make the judgment. Evidence that implicates or supports more than one category tends to induce greater subadditivity than does evidence implicating only a single category. This "cue conflict" interpretation of what Tversky and Koehler (1994) referred to as the enhancement effect was strongly supported in both experiments. (The role of cue frequency is less clear.) When one category is specified for judgment and the alternatives are included implicitly in a residual category, introduction of "mixed" evidence implicating multiple categories is interpreted by participants as supporting differentially the category designated for judgment (cf. Peterson & Pitz, 1988). Koehler et al. (1997) suggest that categories or hypotheses included implicitly in the residual do not utilize the support available from the evidence as efficiently as does the specified, focal hypothesis because the way in which the evidence supports the specified hypothesis is more readily apparent than is the way in which it supports its negation through the alternatives included in the residual.

Acknowledgments

Experiment 1 was conducted during a postdoctoral visit at the Medical Research Council Applied Psychology Unit in Cambridge, England, 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 am grateful to Alan Baddeley and Nigel Harvey for acting as hosts during my visits to these two institutions, respectively. Experiment 3 was conducted at the University of Waterloo, and was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (OGP 0183792). I am grateful to Stephen Lewandowsky, Robert Nosofsky, David Shanks, and Amos Tversky for their comments on an earlier version of this article.

References

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

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

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

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 948*, 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.

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.

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

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.

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.

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

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

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