The Journal of Risk and Uncertainty, 22:3; 227–249, 2001
© 2001 Kluwer Academic Publishers. Manufactured in The Netherlands.

The Relation Between Probability and Evidence Judgment: An Extension of Support Theory*†

LORRAINE CHEN IDSON AND DAVID H. KRANTZ
Columbia University

DANIEL OSHERSON
Rice University

NICOLAO BONINI
University of Trento

Abstract

We propose a theory that relates perceived evidence to numerical probability judgment. The most successful prior account of this relation is *Support Theory*, advanced in Tversky and Koehler (1994). Support Theory, however, implies additive probability estimates for binary partitions. In contrast, superadditivity has been documented in Macchi, Osherson, and Krantz (1999), and both sub- and superadditivity appear in the experiments reported here. Nonadditivity suggests asymmetry in the processing of focal and nonfocal hypotheses, even within binary partitions. We extend Support Theory by revising its basic equation to allow such asymmetry, and compare the two equations' ability to predict numerical assessments of probability from scaled estimates of evidence for and against a given proposition. Both between- and within-subject experimental designs are employed for this purpose. We find that the revised equation is more accurate than the original Support Theory equation. The implications of asymmetric processing on qualitative assessments of chance are also briefly discussed.

Keywords: probability judgment, subaddivity, evidence judgment

1. Introduction

Degrees of belief are implicit in most decisions whose outcomes depend on uncertain events. In quantitative theories of decision making—e.g., Subjective Expected Utility Theory (first axiomatized by Savage, 1954) or Cumulative Prospect Theory (Tversky

*Research in this paper was partly supported by NSF grants SBR-9818849 to D. H. Krantz and IIS-9978135 to D. Osherson.

†We thank Lyle Brenner and Elke Weber for trenchant comments on an earlier draft. Corresponding author: Lorraine Chen Idson, Department of Psychology, Columbia University, New York, NY 10027. E-mail: lorraine @psych.columbia.edu.

and Kahneman, 1992)—degrees of belief are related to decision weights, which can be numerically estimated from the options selected in an appropriately designed set of choices. Depending on the theory, estimated degrees of belief may have some or all of the usual properties of mathematical probability.

People also communicate their degrees of belief in more direct fashion, using a variety of linguistic devices, including judgments of numerical probability (e.g., percentage chance for an uncertain event). Such explicit estimates of chance are often collected systematically and viewed as an approximation to the degrees of belief implicitly involved in decisions (see, e.g., Wallsten, 1971; Morgan and Henrion, 1990; Kleindorfer, Kunreuther, and Schoemaker, 1993; Fox, 1999).

Since numerical probability judgments are easy to obtain, it is important to understand their underlying psychophysics. Biases in judgment can then be diagnosed and perhaps corrected, leading to coherent probabilities derived from subjective judgments (for discussion, see Osherson et al., 1997; Osherson et al., in press).

The well-known work of Kahneman and Tversky (e.g., Tversky & Kahneman, 1974) established that numerical probability judgments are often based on heuristics that produce serious biases. Above all, this work showed that probability judgment is usually based on processes that overlook the extension, or set of exemplars, of a category or event. Often, judgment depends instead on properties that describe the judged categories. Building on the latter insight, Tversky and his collaborators recently introduced a simple and general theory of probability judgment. The theory is motivated by a striking fact about numerical measures of belief (obtained either implicitly from analysis of decisions, or explicitly from estimates of chances). Such measures are often *subadditive:* the numerical value attached to a disjunction of mutually exclusive events is smaller than the sum of the values attached separately to the disjuncts. Tversky and his collaborators suggested that subadditivity arises because the description of a disjunction of two propositions A and B usually brings to mind less evidence than the total evidence obtained when A and B are considered separately.

These ideas were expressed in *Support Theory*, first formulated by Tversky and Koehler (1994) and Rottenstreich and Tversky (1997), then deftly elaborated by Brenner and Koehler (1999). Support Theory distinguishes between logical propositions and their descriptions. Alternative descriptions of the same proposition may bring to mind distinct pieces of evidence; degree of belief is then derived from an evaluation of the evidence favoring the proposition that comes to mind and of the corresponding evidence favoring the alternative or contrary proposition.

Perhaps the primary contributions of Support Theory are its emphases on descriptions and on evaluations of evidence as psychologically fundamental. The theory also includes an equation linking support (evaluation of evidence) to probability judgment. At the time Support Theory was formulated, the existing data (Tversky and Fox, 1995; Tversky and Koehler, 1994; Wallsten, Budescu, and Zwick, 1992) suggested that subadditivity is characteristic of partitions involving three or more mutually exclusive events, but that judgments for binary partitions, where A and B are viewed as mutually exclusive, are additive. The Support Theory equation was designed to accommodate these data and hence is intrinsically symmetric. As discussed in more detail below, this symmetry leads to a prediction of *additivity:* the separate probability judgments sum to 1.

Subsequently, however, violations of additivity have been found for binary partitions. Brenner and Rottenstreich (1999) demonstrated superadditivity for binary partitions in which one of the alternatives is itself a disjunction, and Macchi, Osherson, and Krantz (1999) demonstrated superadditivity for binary partitions where respondents had little knowledge of either alternative. Under the latter conditions, both propositions A and B led to low probability judgments when they were the focal proposition, resulting in probabilities that summed to less than 1. We show below that this can be accounted for by asymmetric processing of a focal proposition and its contrary. Indeed, Macchi, Osherson, and Krantz (1999) found that when questions were reworded to call attention to both A and B, superadditivity disappeared. Such results indicate the need for an extension of Support Theory that introduces a possible asymmetry between the focal proposition and its implicit contrary.

The present work proposes a revision of the basic equation of Support Theory that allows for such asymmetric processing. The revised equation captures the idea that if the judge reflects on the evidence for a proposition but not for its contrary, then the probability estimate will be determined by the perceived amount of evidence for the proposition relative to some constant, which we denote by K. Possibly K may be interpreted as a default value of contrary evidence strength, subject to influence by context and frame, or it may be simply a normalization constant used to convert the open-ended evidence scale into a probability scale. The revised equation also contains a parameter (denoted λ) that reflects the degree to which the judge focuses on the evidence for the contrary proposition, i.e., the degree of symmetric processing. The more symmetric the processing, the closer the revised equation resembles the Support Theory equation. By allowing for asymmetric processing, the revised equation can account for both sub- and superadditivity for binary partitions.

To compare the revised equation with Support Theory, we obtained both probability judgments and evidence judgments for a variety of propositions, in a laboratory setting. We tested the fit of both equations in predicting the probability judgments from the judgments of evidence for and against the corresponding propositions. In several experiments, using both between and within-subject designs, the revised equation produced more accurate predictions than the Support Theory equation.

Our results sustain the idea of Tversky and his collaborators that judgments of evidence strength or support underlie numerical judgments of probability. The data also deepen our appreciation of the complexity of evidence judgments themselves. In particular, they suggest asymmetric processing of evidence in favor versus contrary to a given proposition, and they reveal an unexpected covariance structure for these types of judgment.

The rest of the article proceeds as follows. The first section recapitulates the equations of Support Theory, analyzes the relationship of the terms in these equations to probability judgments for binary partitions, and presents our proposed extension. We then report three studies. Experiment 1 demonstrates that both superadditivity and subadditivity occur in probability judgments for binary partitions. Experiment 2 repeats the demonstration with a larger set of items and explores the relation between probability judgments and evidence judgments. In this study, separate groups of respondents are used for the probability- and evidence-judgment tasks, so the extension of Support Theory is tested only for the relations between group mean judgments. Experiment 3 obtains both probability and evidence judgments from the same respondents, thus permitting within-respondent tests. Finally, we discuss the importance of non-numerical estimates of chance and consider the extent to which the evidence-based theory developed here can be generalized to this setting.

2. Theory

2.1. *Support Theory and binary partitions*

Support Theory distinguishes between logical propositions and their descriptions. Descriptions (including a framing context) can strongly influence the recruitment of evidence for and against the proposition, in both type and amount. Given descriptions A, B of two mutually exclusive propositions, let $P(A, B)$ be the judged probability of A given that exactly one of A, B is true. Let $s(A)$ be the amount of evidence (support) perceived for A, and similarly for $s(B)$. Then Support Theory asserts

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

Immediately from (1) we obtain

Binary Additivity:

$$P(A, B) + P(B, A) = \frac{s(A)}{s(A) + s(B)} + \frac{s(B)}{s(B) + s(A)} = 1. \tag{2}$$

If A is logically equivalent to the exclusive disjunction $A_1 \vee \cdots \vee A_n$, then Support Theory also postulates:

$$s(A) \leq s(A_1) + \cdots + s(A_n), \tag{3}$$

with the inequality often strict.[1] Suppose now that $A_1 \cdots A_n$ is a logical partition (the A_i's are both exclusive and exhaustive). Then $\neg A_i$ is logically equivalent to $A_1 \vee \cdots \vee A_{i-1} \vee A_{i+1} \vee \cdots \vee A_n$, the disjunction of the other $n - 1$ statements. It follows from (1),

(3) that

$$\begin{aligned}\sum_i P(A_i, \neg A_i)& \\ &= \sum_i \frac{s(A_i)}{s(A_i) + s(\neg A_i)} \\ &\geq \sum_i \frac{s(A_i)}{s(A_i) + s(A_1) + \cdots + s(A_{i-1}) + s(A_{i+1}) + \cdots + s(A_n)} \\ &= \sum_i \frac{s(A_i)}{s(A_1) + \cdots + s(A_n)} = 1. \qquad (4)\end{aligned}$$

In order to interpret (4) as a claim about events considered in isolation, we take $P(A, \neg A)$ to be the judged probability $P(A)$ of A. Thus, we assume that judges assign the same probability to the truth of A as they do to the truth of A given either A or $\neg A$. This substantive but reasonable assumption underlies the discussion in Tversky and Koehler (1994), and is adopted explicitly in Fox (1999, p. 174). Replacing $P(A_i, \neg A_i)$ with $P(A_i)$ in (4) yields the prediction:

> *Subadditivity.* For any logical partition $A_1 \cdots A_n$, $\sum_i P(A_i) \geq 1$, with the inequality often strict. (5)

As noted in the introduction, Support Theory was designed to explain additivity in binary partitions. So let us now show that (4) is compatible with (2), binary additivity. If the logical partition in (5) is just $A, \neg A$, then $\neg A$ is equivalent to no disjunction other than the degenerate one consisting of $\neg A$ itself. So (4) becomes

$$\begin{aligned}\sum_i P(A_i, \neg A_i) &= P(A, \neg A) + P(\neg A, \neg\neg A) = P(A, \neg A) + P(\neg A, A) \\ &= \frac{s(A)}{s(A) + s(\neg A)} + \frac{s(\neg A)}{s(\neg A) + s(A)} = 1, \qquad (6)\end{aligned}$$

where we assume no difference in the respondent's estimates of $P(\neg A, \neg\neg A)$ and $P(\neg A, A)$. Relying again on the identification of $P(A)$ with $P(A, \neg A)$, and $P(\neg A)$ with $P(\neg A, A)$, (6) yields:

> For the logical partition $A, \neg A$, $P(A) + P(\neg A) = 1$. (7)

Hence, (4) is consistent with the claim of binary additivity in (2), when B in the latter equation is equivalent to $\neg A$. In contrast, for the nontrivial case $A_1 \cdots A_n$, where $n > 2$, the inequality in (3) can be strict, yielding strict inequality in (4), hence the subadditivity prediction (5).[2]

The studies reviewed in Tversky and Koehler (1994) and Rottenstreich and Tversky (1997) sustain (2), (5), and other aspects of Support Theory. With the aid of supplementary assumptions, Brenner and Koehler (1999) derive further predictions from (1), (3), and confirm them experimentally. Macchi, Osherson, and Krantz (1999), however, describe binary partitions A, B such that $P(A) + P(B) < 1$ (*superadditivity*), contrary to the predictions of Support Theory. For example, one group of Italian undergraduates was asked for the probability that the Duomo in Milan is taller than Notre Dame in Paris, whereas another group was asked for the probability that Notre Dame is taller than the Duomo. Both groups were informed that the heights are not identical. Despite many answers of 0.5, the sum of the average answers for the two groups was only 0.72. The same pattern was observed for the other items in the study, across three replications with different respondents. (Answers of 0.5 can be interpreted as culturally sanctioned expressions of complete ignorance—see Fischhoff and Bruine de Bruin, 1999.) Brenner and Rottenstreich (1999) also demonstrated superadditivity.

To summarize, the literature shows evidence for both additivity and superadditivity for binary partitions. In the partitions used by Macchi, Osherson, and Krantz to demonstrate superadditivity, it is plausible that little evidence comes to mind for any member of the partition. On the other hand, subadditivity is explained in Support Theory by the fact that substantial evidence can be brought to mind for each partition element taken alone. This suggests that subadditivity depends on the availability of substantial evidence in favor of partition elements whereas superadditivity depends on its dearth. Perhaps both can be found for binary partitions; here, deviations in either direction from additivity would depend on asymmetry of processing of the focal proposition and its contrary.

2.2. Extended Support Theory (EST)

In this subsection we advance an extension of Support Theory that incorporates asymmetry and permits both sub- and superadditivity for binary partitions.

We consider binary partitions, such as the following:

> A: The Indian rhinoceros will become extinct before the Northern Pacific sea-horse.
> B: The Northern Pacific sea-horse will become extinct before the Indian rhinoceros. (8)

Notice that B is equivalent to the complement of A, and vice versa.[3] If asked for the probability $P(A)$ of A in isolation, the judge might reflect not just on the evidence in favor of A, but also on the evidence against it. This is tantamount to considering the evidence for $\neg A$ which, as just observed, is equivalent to considering the evidence for B. The judge's answer will thus have the form $P(A, B)$. In this case we follow Equation (1) of Support Theory, and take the assessed probability to be $s(A)/[s(A) + s(B)]$, provided that at least one of these is greater than 0. On the other hand, the judge might not reflect upon the evidence against A. We then assume that $P(A)$ is derived from $s(A)$ alone. In Support Theory, however, $s(A)$ is treated as an unbounded ratio scale; and in experiments, direct judgments of evidence strength are usually made on a scale that is not bounded by 100% (Edwards, Lindman, and Phillips, 1965; Briggs and Krantz, 1992). A simple mapping of $s(A)$ to a scale bounded by 100% is given by $s(A)/[s(A) + K]$, where K is a constant. The latter constant may be interpreted as a default value of contrary evidence strength, replacing $s(B)$ in situations where the judge does not reflect on B. Such a default value might possibly be manipulated by changing context, e.g., by introducing a series of propositions that have very strong contrary evidence, or the reverse. Alternatively, K could be interpreted simply as a device used to map an open-ended continuum of evidence evaluation into the finite interval [0,1], and thus may depend on the scale used to evaluate evidence. Let λ denote the probability of attending to evidence against A when estimating $P(A)$. We assume that λ depends on the judge and on factors that influence attention to the contrary proposition.

In place of $s(A)$ and $s(\neg A)$ from Support Theory, we let $\mathtt{for}(A)$ be the perceived evidence in support of A, and $\mathtt{against}(A)$ be the perceived evidence contrary to A (if any comes to mind). Our hypothesis may thus be stated as follows.

Extended Support Theory: Fix a judge and a broad class of statements drawn from a common domain. Then there is $\lambda \in [0, 1]$ and positive constant K such that whenever either $\mathtt{for}(A) > 0$ or $\mathtt{against}(A) > 0$:

$$P(A) = \lambda \times \frac{\mathtt{for(A)}}{\mathtt{for(A)} + \mathtt{against}(A)} + (1 - \lambda) \times \frac{\mathtt{for(A)}}{\mathtt{for(A)} + K}. \tag{9}$$

Observe that (9) is compatible with both sub- and superadditive judgment. For λ below unity, subadditivity will tend to occur when the judge perceives substantial evidence in favor of both members A, B of a binary partition (i.e., both $\mathtt{for}(A)$ and $\mathtt{for}(B)$ are high). Superadditivity will tend to occur when little evidence is perceived. Strict additivity is also possible, either because λ is unity, or because K approximates the perceived amount of evidence against A, B. By contrast, Support Theory predicts additivity for all binary partitions.

If evidence neither in favor of A nor against it come to the judge's mind, then both $\mathtt{for}$(A) and $\mathtt{against}$(A) are zero, so $\mathtt{for}(A)/(\mathtt{for}(A) + \mathtt{against}(A))$ is undefined. In this case, EST makes no prediction. Total ignorance about a proposition is often expressed by assigning it probability 0.5 (Fischhoff and Bruine de Bruin, 1999), so we may supplement the theory by predicting an assessed probability of 0.5 when both $\mathtt{for}$(A) and $\mathtt{against}$(A) are zero.

The theory formulated in the present article represents only one of several approaches to modeling the asymmetric processing of a focal proposition and its contrary. A different relation between numerical probability and evidence has been suggested to us by Brenner and Rottenstreich. It can be expressed by the following alternative to equation (9)

$$P(A) = \mathtt{for}(A)/[\mathtt{for}(A) + \lambda \times \mathtt{against}(A) + (1 - \lambda) \times K].$$

This equation fits our data a little less well than EST (as we have formulated it). The alternative equation may nonetheless throw light on aspects of numerical probability assessment not explored here.

3. Experiment 1

EST is motivated by the dual phenomena of sub- and superadditive judgment. Subadditivity was originally documented by Fischhoff, Slovic, and Lichtenstein (1978), and confirmed in subsequent studies (e.g., Russo and Kolzow, 1994; Tversky and Koehler, 1994).

Superadditivity was described by Cohen, Dearnaley, and Hansel (1956), then emphasized by Brenner and Rottenstreich (1999) and by Macchi, Osherson, and Krantz (1999). According to (9), both sub- and superadditivity should be obtainable in the same group of respondents, by varying the amount of evidence that comes to mind when estimating probabilities. Because published demonstrations of superadditivity are few, it seemed important to exhibit the phenomenon once more and to relate it to respondents' domain knowledge. This was the goal of the present experiment.

3.1. Method

Six binary partitions were created. Three were intended to evoke substantial evidence in the minds of our (Italian) judges, the others were intended to evoke little evidence. Each partition gives rise to symmetrical questions, such as the following:

> Assuming that Morocco and Kenya make it to the next World Cup Finals, what is the probability that Morocco wins?
>
> Assuming that Morocco and Kenya make it to the next World Cup Finals, what is the probability that Kenya wins?

The former question will be called "version *A*," the later "version *B*." Version *A* questions from each partition are provided in Appendix A.

Eighty-one undergraduates from the University of Cagliari (Sardinia, Italy) estimated the probability of the *A* versions; eighty estimated the *B* versions. Students were assigned randomly to the two groups, and run in a classroom setting. Questions were presented via booklets with individually randomized order.

3.2. Results

Mean probability estimates for the two versions of the six partitions are shown in Table 1. For each partition we computed the sum of these mean estimates for the two versions over all respondents, and tested its difference from 1.00 via a two-tailed t-test. Four of the partitions showed deviations from additivity in the predicted directions, with attained significance levels as shown in the last column of the table. This included two high-knowledge items (*Election winner* and *Italian soccer*), which showed subadditivity, and two low-knowledge items (*Traffic fatalities* and *Record broken*), which showed superadditivity. The other two items showed non-significant deviations in the direction opposite to that predicted.

Table 1. Experiment 1. Mean probability estimates for propositions A ($N = 81$) and B ($N = 80$) for the six partitions

	Means			
Propositions	*A*	*B*	Sum: 90% C.I.	Attained significance
Designed high knowledge				
Election winner: Olive Branch vs. Liberty	0.608	0.446	1.054 ± 0.052	0.09
Italian soccer: Inter vs. Juventus	0.498	0.581	1.079 ± 0.058	0.02
Olympic site: Italy vs. South Africa	0.566	0.413	0.979 ± 0.054	0.52
Designed low knowledge				
Traffic fatalities: 1998 vs. 1997	0.418	0.422	0.840 ± 0.062	0.00003
Record broken: high jump vs. 10 km walk	0.428	0.477	0.905 ± 0.060	0.01
World Cup winner: Morocco vs. Kenya	0.509	0.521	1.030 ± 0.045	0.27

Also given are 90%-confidence intervals for the sum of the means and attained significance levels for two-tailed tests of the hypothesis: sum $= 1$.

3.3. Discussion

These results demonstrate superadditivity for some items designed to evoke little evidence, in accord with the findings of Macchi, Osherson, and Krantz (1999). Subadditivity also occurs for some items intended to evoke much evidence. The results are not found for every item, however, and it would clearly be desirable to have independent information about the degree of knowledge actually available in the respondent population for these items. The next experiment therefore obtained both probability judgments and direct evidence judgments, from two groups or respondents sampled from the same population.

4. Experiment 2

The previous experiment, as well as those reported in Macchi, Osherson, and Krantz (1999), were performed with Italian participants. We devised 16 new binary partitions for use with American undergraduates. They are also presented in Appendix A. Again, it was intended that the partitions be variable in how much evidence their members bring to the minds of undergraduates. For example, we expected more evidence to result from thinking about crime in New York City than about unemployment in Finland and Estonia (see items 5 and 14 in Appendix A).[4] Some respondents provided evidence-ratings for the first proposition (version A) in each partition, the others for version B, and similarly for probability.

4.1. Method

Four groups of respondents were created by random selection from a pool of 136 Columbia University undergraduate volunteers. Thirty-three respondents estimated the probability (on a percentage scale, 0–100%) of the *A* statements in each partition. Another thirty-three estimated probabilities for the *B* statements. Thirty-five additional respondents rated the *A* statements in terms of the amount of evidence in its favor and the amount of evidence against it. Ratings of positive and negative evidence used the six categories are displayed in Table 2.[5] Respondents were required to write their evaluation (chosen from these 6 categories) on two blank lines, one labeled "evidence for," the other labeled "evidence against." Another thirty-five respondents rated the *B* statements in the same way. We selected one ordering of the items at random and used it for half of the respondents, the reverse order for the remaining respondents.

The identical procedure was carried out with 154 undergraduates at Rice University. Forty-three respondents estimated probabilities for *A* statements, and 43 for *B* statements. Another 34 students rated the *A* statements for evidence in favor and against. Still another group of 34 students rated the *B* statements.

4.2. Results

Probability judgments: sub- and superadditivity. Figure 1 shows the principal results concerning probability judgments. Points are plotted using item numbers (1–16) in Appendix A. Columbia data are plotted on the abscissa, Rice data on the ordinate. The plotted values are sums of the mean probability judgments for the *A* and *B* statements for each partition. Perfect additivity would put all points at (1, 1). Our predictions, on the basis of prior guesses concerning respondents' high or low knowledge, were that

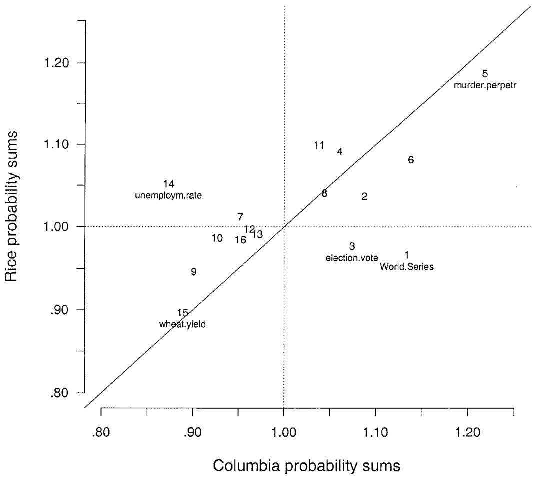

Figure 1. Mean probability sums for Experiment 2. The integers refer to the partitions of Appendix A. Each point has abscissa equal to the sum of mean probability judgments for that partition at Columbia, and ordinate equal to the corresponding sum at Rice. Items falling in the upper right quadrant were subadditive for both groups, those in the lower left were superadditive for both.

problems 1–8 would be subadditive, falling into the upper right quadrant and partitions 9–16 would be superadditive, falling into the lower left quadrant.

The results for six problems fall into the upper right quadrant, hence, they were subadditive for both groups of respondents. Five of the six were predicted to be so. Six other problems fall into the lower left quadrant, hence showing superadditivity for both groups. All six were predicted to be so.

Points falling on or near the 45° line are good replications across the two groups. The two problems that were highly subadditive or superadditive in both groups are labeled as to content, as are the three problems where the two groups gave sharply different results. For one of these latter three, however, the result obtained makes sense *post hoc*. The World Series problem asked about the Yankees and Mets, two New York teams. This acted as predicted (high knowledge, subadditive) at Columbia, in New York, but not at Rice, in Texas.

Separate t-tests suggest that subadditivity was statistically reliable for 3 or 4 of the partitions at Columbia (1, 3, 5 and 6). Problem 2 (women in the priesthood by 2050) showed too much variability to permit a reliable conclusion about subadditivity. Separate t-tests also suggest reliable superadditivity for 3 partitions at Columbia (9, 14 and 15). In general, the deviations from additivity were smaller at Rice, and the only ones that were statistically reliable were partitions 4 and 5 (subadditive) and 15 (superadditive).

The overall pattern of results supports the general thesis that low knowledge leads to superadditivity, high knowledge to subadditivity. For a sharper test of EST we compare the respondents' probability for a given proposition X with the predictions that result from inserting $\mathtt{for}(X)$ and $\mathtt{against}(X)$ into Equation (9). (X can be either the A or B propositions in any of the 16 partitions shown in Appendix A.) Since probability and evidence judgments came from different sets of respondents, only the relationship between the means will be explored. Prior to reporting the results of this analysis, however, we make some comments on the evidence judgments themselves. First we consider $\mathtt{for}(X)$ alone, then discuss $\mathtt{for}(X)$ and $\mathtt{against}(X)$ together.

Judgments of positive evidence. We coded the six evidence categories (Table 2) as 0–5. The averages of *evidence-for* judgments for both the A and B versions of the partitions are plotted in Figure 2. As in Figure 1, the 16 items from Appendix A are used as plotting symbols, with the mean for Columbia respondents as the abscissa, and the Rice mean as the ordinate.

Table 2. Rating categories for positive and negative evidence, with numerical codes used for analysis

Numerical code	Evidence for	Evidence against
0	Little or no evidence	Little or no evidence
1	Weak evidence in favor	Weak evidence against
2	Moderate evidence in favor	Moderate evidence against
3	Strong evidence in favor	Strong evidence against
4	Very strong evidence in favor	Very strong evidence against
5	Overwhelming positive proof	Overwhelming contrary proof

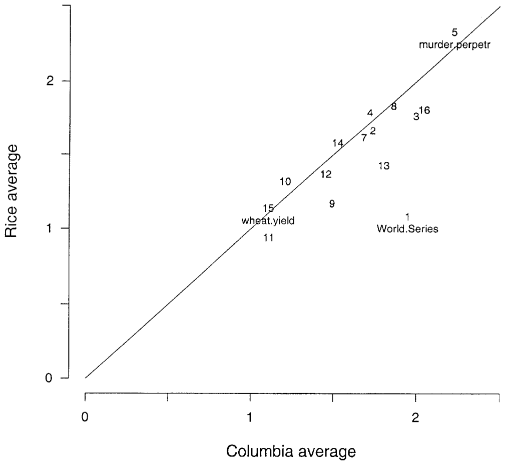

Figure 2. Evidence-for judgments in Experiment 2. The plotted numerals identify (A, B) partitions in Appendix A. The abscissa value is the average Columbia *evidence-for* judgment for the partition, including all respondents and both versions (A and B). The ordinate value is the corresponding Rice average.

It is striking that the evidence judgments lie near the 45° line, except for the World Series item. This shows close replicability of the two sets of results, apart from an understandable departure in the World Series partition. The most extreme subadditive item from Figure 1 (murder perpetrator) is revealed in Figure 2 to bring abundant positive evidence to mind. The most extreme superadditive item (wheat yield) is similarly confirmed as failing to recruit such evidence.

Bivariate structure in evidence judgments. When we consider judgments of both $\mathtt{for}(X)$ and $\mathtt{against}(X)$, a different picture emerges. For superadditive items, many evidence-judgment pairs are (0, 0), that is, respondents use the bottom of the evidence-judgment scale for both questions, *for* and *against*. For most of these same items, however, there are also quite a few judgments of form (j, j), with $j > 0$, i.e., many respondents judge that evidence is equal for and against, but not minimal. Relatively fewer respondents give judgments of form $(j, 0)$ or $(0, j)(j > 0)$ for these items. Because of the large number of (0, 0) and (j, j) judgments for superadditive items, such items exhibit a positive correlation between $\mathtt{for}(X)$ and $\mathtt{against}(X)$, across respondents. For the subadditive items, on the other hand, there are many fewer (0, 0) judgments, somewhat fewer (j, j) judgments, and many more polarized judgments of form $(j, 0)$ or $(0, j)$. Thus, the correlation between $\mathtt{for}(X)$ and $\mathtt{against}(X)$ across respondents tends to be near zero or even negative for subadditive items.

We address this phenomenon more precisely in the following analysis. For each of the thirty-two propositions X arising from the partitions of Appendix A, we compared (a) the correlation of $\mathtt{for}(X)$ with $\mathtt{against}(X)$, and (b) the mean of $\mathtt{for}(X)$ and $\mathtt{against}(X)$. These comparisons were carried out for both the Columbia and Rice replications of the present experiment, and also for the evidence-rating portion of a third replication carried out at Columbia (and reported below as Experiment 3). These 96 comparisons (involving 16 partitions × 2 versions × 3 replications) are plotted as small circles in Figure 3. (There are no important differences among the separate scatterplots for versions or replications.) The numerals in the figure show the 16 points obtained by averaging correlations and means over the two versions and three replications.

The relationship shown in Figure 3 was unexpected, and too strong to ignore. It is not related to the questions addressed in this article, however, so it is not explored further.

Probability related to evidence: mean judgments. Let us now consider the relationship between mean evidence judgments and mean probability judgments. For this analysis, we considered the thirty-two statements resulting from the A and B versions of our 16 partitions, and we attempted to fit equation (9) after coding the evidence judgments as 0–5 (as before).

For the Columbia and Rice data separately, we fit the λ and K parameters of equation (9), minimizing the sum of squared deviations between the 32 mean probability estimates $P(X)$ and the predictions from (9) based on the mean evidence judgments $\mathtt{for}(X)$ and $\mathtt{against}(X)$. The fits were obtained by a 2-dimensional grid search, with steps of 0.25 for K and 0.04 for λ.

Part (a) of Figure 4 shows the relation between mean probability judgments and the predictions from the mean evidence judgments. Since the Columbia and Rice parameter values were similar, average values were used for the predictions displayed here. These predictions are quite accurate for all 32 propositions (A and B versions) and are not

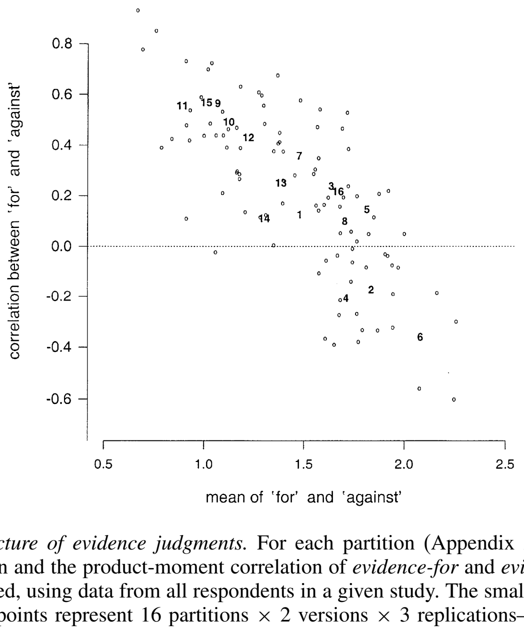

Figure 3. Bivariate structure of evidence judgments. For each partition (Appendix A) and each proposition (version A or B) the mean and the product-moment correlation of *evidence-for* and *evidence-against* judgments (scaled 0–5) was computed, using data from all respondents in a given study. The small circles plot these means and correlations: the 96 points represent 16 partitions × 2 versions × 3 replications—the Columbia and Rice studies reported in this section and a replication at Columbia (reported in experiment 3). The numerals show the means of the six points for each partition.

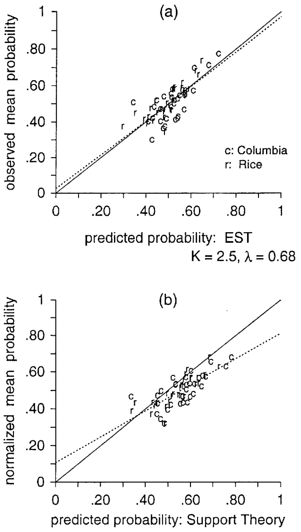

Figure 4. (a) *Prediction of mean probability judgments from mean evidence judgments using equation* (9). The 64 points include all 32 *A* and *B* statements (Appendix A) from both Columbia and Rice subgroups. The fit involved 2 free parameters, as shown. (b) *Prediction of normalized probability judgments from Support Theory.* Mean probability judgments for each *A* and *B* statement were normalized by dividing by their sum. This involves 32 implicit parameters (since normalization was done separately for Rice and Columbia); there are no free parameters in Support Theory *per se.*

appreciably different for Columbia and Rice groups, though the residual variance is higher for the Columbia data. The figure plots all 64 (predicted, observed) pairs across the two replications of the experiment. The solid line has intercept 0 and slope 1; it differs little from the regression line (dashed).

The fit of Support Theory ($\lambda = 1$) to these data is much inferior: the sum of squared residuals is about double, for each group. Importantly, only a small part of the superiority of equation (9) is attributable to deviations from additivity. If the *A* and *B* mean probability judgments are normalized to sum to 1 (dividing each by their sum) the sum of squared deviations is reduced only slightly. This fit of Support Theory is shown in part (b) of Figure 4. In part (b), the least-squares line differs substantially from the line with slope 1 and 0 intercept.

We tentatively conclude that the relationship between evidence and probability judgment is fairly well described by equation (9) of EST. Moreover, the theory appears to be superior in this respect to Support Theory. Although equation (9) was motivated by deviations from additivity, such deviations seem to be only a small part of the reason why Support Theory fails as a detailed description of this relationship.

Both Support Theory and EST are meant to apply to individual judges rather than to averages. The theories are therefore better compared when both evidence and probability assessments are drawn from the same respondent. Data of this kind were obtained in the experiment reported next.

5. Experiment 3

Within-respondent tests of Support Theory have been carried out by Fox (1999). These experiments are extremely illuminating but limited by the fact that judges were highly knowledgeable about the domains in question (e.g., professional basketball). They were thus likely to bring ample evidence to bear on each probability estimate. The present study employed the partitions of Experiment 2, which were designed to vary the amount of evidence that comes to the minds of our undergraduate judges.

5.1. *Method*

Forty students at the Columbia University summer session (all native English speakers) provided both probabilities and evidence ratings. Eighteen participants considered only the *A* propositions of each partition. For these propositions they completed both the probability and evidence-rating tasks described in Experiment 2. The same ordering of partitions was used as in Experiment 2 (with half the respondents receiving the reversed ordering). After completing probability judgments for the first four propositions, respondents made evidence judgments for and against those same items, then, after an unrelated intervening task, they continued with the next 4 propositions, etc. Likewise, twenty-two respondents gave probability and evidence judgments for the *B* propositions of the partitions.

5.2. *Results*

Mean judgments. The mean probability judgments, and the mean evidence judgments (coding the categories as 0–5) were similar to those of Experiment 2. The mean probability sums for versions *A* and *B* correlated 0.46 with the Columbia mean sums and 0.71 with the Rice mean sums from Experiment 2. For comparison, the correlation between the Columbia and Rice sums, shown in Figure 1, was 0.63. The evidence judgments correlated much more highly: the means of the evidence-for judgments correlate, across partitions, about 0.88 with the Columbia data and about 0.77 with the Rice data from Experiment 2.[6] (The Columbia–Rice correlation, shown in Figure 2, was 0.85.) Correlations across experiments are even higher for mean probability or evidence judgments than for probability or evidence sums; this is because the sums have a more restricted range.

As in Experiment 1, our predictions about sub- and superadditivity were only partly fulfilled. Of the eight items designed to be low knowledge, six of them were in fact superadditive, as shown by sums of means. Moreover, the item-by-item replication of the earlier Columbia data was quite close except for one deviant item. The responses to the items designed to be high knowledge were rather different from Experiment 2, however. Only the most extreme item (partition 5, “murder perpetrator”) replicated well, and only four of the items actually showed subadditivity. There were no outright reversals, however, in which items predicted to deviate from additivity in one direction deviated sharply in the opposite direction.

Our failure to predict exactly which items are sub- or superadditive may be explainable by the idea that deviations from additivity require asymmetric processing. For some items, consideration of evidence for may often lead to consideration of evidence against, thus bringing the item closer to additivity.

Individual judgments. The main analysis for Experiment 3 focusses on the relationship between evidence and probability judgments within respondents. As before, we coded evidence on a 0–5 scale. This resulted in $\mathtt{for}(X) = \mathtt{against}\ (X) = 0$ for some of the propositions evaluated by some of the respondents. Equation (9) is undefined in such a case. We thus applied equation (9) only to judgments where the respondent indicated some information about the matter at hand, by an evidence judgment other than (0, 0).

As an aside, we note that the judgment of (0, 0) likely expressed no knowledge about the material at hand. This judgment pair occured for 28% of the low-knowledge items as compared with 7.5% of the high-knowledge ones. In both cases, it was accompanied over 60% of the time by a probability judgment of 50. The latter value can be assimilated to the conventional use of "50–50" to signal ignorance (Fischhoff and Bruine de Bruin, 1999).

Of the 40 respondents, only 32 provided data that seemed suitable for testing equation (9). Four were dropped because they responded with (0, 0) to 7 or more evidence items, leaving fewer than 10 items for parameter estimation in equation (9). Four additional respondents were dropped because their response patterns showed stereotypy (absence of variability) from one item to the next: for example, one of them gave the evidence judgment of (2, 2) 13 out of 16 times.

We fit equation (9) to the 32 remaining data-sets. They yielded 435 testable predictions, compared to $32 \times 16 = 512$ potential ones. This reduction comes almost entirely from the elimination of (0, 0) evidence judgments (there are also a very small number of missing data for these respondents). To fit the equation, we first allowed the parameters λ and K to vary among respondents. In each case, we used a two-dimensional grid search to minimize the sum of squared deviations between observed and fitted probability judgments. The fit is summarized across all 32 respondents by the analysis-of-variance summarized in Table 3 (part a). In this analysis, we take Support Theory (ST) as a baseline model, for which the total sum of squared deviations (based on 435 testable items for the 32 respondents) is 33.93. By contrast, EST reduces this sum of squared deviations to 22.26, fitting 64 parameters (λ and K for each respondent). The F-ratio of 3.04 strongly rejects ST as a special case of EST. A more detailed analysis of individual differences in fit and individual values of λ and K will be given below.

To probe the value of equation (9) further, we selected 15 respondents for whom Support Theory provides quite a good fit. The selection criterion was that the sum of squared residuals from Support Theory should be less than 1.0 and, in addition, the regression line for a linear fit between the prediction of Support Theory and the observed

Table 3. Comparison of the fit of EST

Source	DF	SS	MS	F
(a) ANOVA table comparing the fit of EST with that of Support Theory for 32 respondents				
Reduction for equation (9)	64	11.67	0.1823	3.04
Residuals from equation (9)	371	22.26	0.0600	
Total (Residuals from ST)	435	33.93		
(b) Comparison of the fits of EST and Support Theory for the 15 respondents with best fit to Support Theory				
Reduction from equation (9)	30	1.633	0.0544	2.04
Residuals from equation (9)	179	4.779	0.0267	
Total (Residuals from ST)	209	6.412		
(c) Comparison of the fit of EST with fixed parameters to the fit of Support Theory, and isolating the variance component for individual differences in parameters, for the 21 respondents with best fit to EST				
Reduction for equation (9) (fixed parameters)	2	2.822	1.421	44.5
Reduction for individual differences	40	2.037	0.051	1.6
Residuals from equation (9)	239	7.638	0.032	
Total (Residuals from ST)	281	12.600		

probability judgment should have slope at least 0.5 for that respondent. For these 15 respondents, the corresponding ANOVA is summarized in Table 3 (part b).

For the sake of symmetry, we likewise selected 21 respondents for whom equation (9) gives a particularly good fit. For these respondents the difference between Support Theory and the best fit of EST was quite large. For this group of respondents, it is also worth exploring the extent of individual differences in the parameters of equation (9). About one-third of these respondents had their best fit with a very low value of λ, while the remainder were scattered over the interval $(0, 1)$. For respondents with $\lambda < 0.5$, estimates of K were scattered between 0 and 3 on the 0–5 scale. (For larger values of λ, estimates of K cannot be very accurate.)

The individual differences just observed do not have much effect on the fit of our theory. This can be seen in the final ANOVA, which compares the fits of Support Theory (as the baseline), equation (9) with fixed parameters ($\lambda = 0.42, K = 2.0$), and equation (9) with variable parameters for these 21 respondents. See Table 3 (part c). Although the difference between Support Theory and EST for this selected group is quite large, the use of individual values of λ and K leads to little improvement in the fit of equation (9). The model with fixed parameters has $r^2 = 0.224$, while the additional 40 individual difference parameters yield an incremental $r^2 = 0.162$.

One might ask whether the individual differences are statistically significant. On the surface, an F test would seem to reject the hypothesis of a null variance component for individual differences [$F(40, 239) = 1.59$, $p \approx .02$]. This statistical claim does not seem worth pressing, however. On the one hand, we have little confidence that F statistics as calculated here would (under the null hypothesis) closely approximate the F distribution with appropriate degrees of freedom. On the other hand, it is obvious in the overall data set that there are substantial individual differences in responses to these tasks, so the null hypothesis has no credibility anyway. To put matters another way, individual differences certainly exist, but sampling variability of individual parameters for equation (9), estimated from 13 or 14 judgments per respondent, (excluding (0, 0) evidence pairs) is too large for reliable identification of such differences.

6. General discussion

In developing Support Theory, Tversky and his collaborators highlighted the relationship between recruitment of evidence and construction of degrees of belief. Except for Fox (1999), however, we do not know of previous studies of the relationship between direct judgments of evidence and degrees of belief or decision weights.

Our results support the thesis that judgments of evidential strength can be used to predict judgments of probability. The predictions of our extended version of Support Theory are reasonably accurate at the level of group means, and also for some individuals. Specifically, the experiments reported above suggest that evidence for a proposition and against it are often processed asymmetrically. The outcome can be super- or subadditivity, even for binary partitions. Thus, the link between evidence and probability has to permit nonadditivity for binary partitions, as equation (9) does.

Asymmetric processing of evidence for and against a stated proposition (compared to its implicit contrary), is not a novel finding. Similar asymmetry has been found in the context of hindsight bias (Christensen-Szalanski and Willham, 1991), in developing arguments for and against risk-taking (Beyth-Marom et al., 1993), and in the study of overconfidence (Koriat, Lichtenstein, and Fischhoff, 1980).[7] The latter study showed that leading judges to focus on evidence against an initial guess is an effective method of reducing overconfidence, producing well-calibrated judgments of numerical probability.

The present experiments also indicate the richness of direct evidence judgments. At the same time, the complicated covariance structure of these judgments (Figure 3) shows that much remains to be learned about the mental representation of evidence strength and about tasks designed to probe it.

Our results suggest that judgments of evidence may be more fundamental than numerical probability judgments, and that the relationship between the two is complex. Some skepticism might thus be warranted concerning the direct use of numerical judgments in calculations of probability or expectation. It seems wiser to treat numerical estimates of chance as *behavioral indicators* of underlying evidence. Since perceived evidence is likely to be scaled differently than probability, incoherence in probability judgment seems the inevitable outcome, and is in fact widely observed (Yates, 1990; Baron, 1994; Osherson, 1995). Thus, if useful probabilities are to be reconstructed from human judgment, some means of approximating them with a coherent distribution may be essential (Osherson et al., 1997; Osherson et al., in press).

Numerical probability judgment, however, is only one of several ways in which people communicate their conviction about a given proposition. We suspect that qualitative expressions for uncertainty provide a better picture of the mental structure of degrees-of-belief than do point probabilities (see Windschitl and Wells, 1996). So it is worth asking whether our results will generalize outside the limited setting of numerical judgment. To address this question, consider deviations from additivity in a qualitative context. Focussing on proposition A may lead someone to categorize it as "rather likely." But if asked to focus on B, the same person may recruit evidence for B and thus also categorize B as "rather likely," even though A and B are viewed as mutually exclusive. Superadditivity, in contrast, seems less likely to arise with qualitative expressions. This is because A can be categorized as "rather unlikely" only if people are paying attention to evidence against A, and it may be difficult to both focus on A yet consider only evidence against it.

Indeed, we have conducted experiments in which college students responded to the partitions of Experiments 2 and 3 using a strictly qualitative response mode (they had to choose a suitable expression of uncertainty from a list). Upon translating the qualitative responses into numerical values, we found ample evidence of subadditivity for binary partitions, but no sign of superadditivity. Superadditivity may therefore be a phenomenon only observable in numerical assessment, where low probability judgments arise through some process described roughly by equation (9).

Thus, while we can claim some success in relating direct evidence judgments to numerical expressions of degree of belief, it is presently unclear how to formulate a unified theory that relates evidence to degree of belief in both qualitative and quantitative expressions.

Appendix A: Propositions used in Experiments 1–3

Version A questions used in Experiment 1

Designed as high-knowledge items.

Election winner. Assuming that the principal parties figuring in the next national election are the "Olive Branch Coalition" and the "Front for Liberty," what is the probability that the Olive Branch Coalition receives more votes?

Italian soccer. Assuming that "Inter" and "Juventus" reach the finals of the next Italian soccer championship, what is the probability that Inter wins?

Olympic site. Assuming that the two finalists for the site of the next Olympic Games are Italy and South Africa, what is the probability that Italy is chosen?

Designed as low-knowledge items.

Traffic fatalities. Assuming that the number of traffic fatalities in France for 1998 is different from the number in 1997, what is the probability that the number is greater in 1998?

Record broken. Assuming that the world record for the high jump is not broken on the same day as the world record for the 10-km walk, that is the probability that the high jump record is broken first?

World Cup winner. Assuming that Morocco and Kenya make it to the next World Cup Finals, what is the probability that Morocco wins?

Partitions for Experiments 2 and 3

Designed as high-knowledge items.

1. Assume that the Yankees and the Mets compete in the 1998 World Series.
   A: The Yankees win.
   B: The Mets win.
2. Assume that in the year 2050, there is a serious discussion in the Vatican Curia about priesthood for women.
   A: The Cardinals in the Curia favor priesthood for women at that time.
   B: The Cardinals in the Curia oppose priesthood for women at that time.
3. Assume that the Democrats and the Republicans are chief contestants in the next national election.
   A: The Democrats get more votes.
   B: The Republicans get more votes.
4. Assume that in the year 2020, there is a vote in the New York State legislature on whether to make sex education compulsory starting in the 4th grade.
   A: The legislature votes in favor of making sex education compulsory starting in the 4th grade at that time.
   B: The legislature votes against making sex education compulsory starting in the 4th grade at that time.
5. *A*: The first murder in New York City in 1999 is committed by someone whom the victim knew.
   B: The first murder in New York City in 1999 is committed by someone whom the victim did not know.
6. *A*: U.S. government expenditures exceed revenues in fiscal year 1999.
   B: U.S. government revenues exceed expenditures in fiscal year 1999.
7. Imagine a student who is very interested in the growth of the Russian economy after the Revolution.
   A: The student is a major in Economics rather than History.
   B: The student is a major in History rather than Economics.
8. *A*: The U.S. will have its first white female President before its first African-American President.
   B: The U.S. will have its first African-American President before its first white female President.

Designed as low-knowledge items.

9. *A*: The fatality rate from traffic accidents in France will be higher in 1998 than in 1997.

B: The fatality rate from traffic accidents in France will be higher in 1997 than in 1998.

10. *A*: The Indian rhinoceros will become extinct before the Northern Pacific sea-horse.
    B: The Northern Pacific sea-horse will become extinct before the Indian rhinoceros.
11. *A*: The world record for the 10-km walk will be broken before the world record for the high jump.
    B: The world record for the high jump will be broken before the world record for the 10-km walk.
12. Imagine that Kevin and James are six-month-old babies who are regular patients at a clinic. Kevin was prenatally exposed to cocaine. James was prenatally exposed to alcohol. Both develop an allergy to penicillin.
    A: Kevin's allergy turns out to be more severe than James'.
    B: James' allergy turns out to be more severe than Kevin's.
13. Imagine that Albert and Ned are fly fishing in the same stream. Albert is using nightcrawlers for bait and Ned is using a store-bought fishing lure.
    A: Albert catches the first fish.
    B: Ned catches the first fish.
14. *A*: In 1998 the unemployment rate in Finland will be higher than in Estonia.
    B: In 1998 the unemployment rate in Estonia will be higher than in Finland.
15. Imagine that a farmer plants wheat in two adjacent fields. In the previous year, he had planted one of these fields with potatoes and the other with cabbage.
    A: The field that previously had potatoes produces more wheat per acre than the field which had cabbage.
    B: The field that previously had cabbage produces more wheat per acre than the field which had potatoes.
16. *A*: A vaccine for AIDS will be discovered before a procedure for regenerating nerves in the spinal cord.
    B: A procedure for regenerating nerves in the spinal cord will be discovered before a vaccine for AIDS.

Notes

[1] The symbol $\vee$ is used to form disjunctions, and may be read as "or." In this paper, disjunctions will only be written between exclusive propositions. We use $\neg A$ to denote the negation of the claim that A occurs.

[2] As noted earlier, the present paper bears on subjective probability only. Subadditivity in frequency estimates is reported in Fiedler and Armbruster (1994), Tversky and Koehler (1994), and Mulford and Dawes (1999). Fiedler and Armbruster (1994) explained their findings in terms of regression to the mean by estimates of low-frequency events. Mulford and Dawes (1999) observed that this interpretation is at variance with their own data on the severity of subadditivity as a function of event frequency.

[3] We view the proposition that there is perfectly simultaneous extinction as having zero or negligible probability; our interviews with respondents in these tasks suggest they likewise ignore such possibilities. We also presume that the respondents accept the implied conditioning event for each partition—in this example, that both will eventually become extinct. We saw no formal or informal evidence that respondents resisted such implicit conditionals, though of course it might be possible to design items that would produce resistance in substantial numbers of respondents. Finally, we do not think that there is appreciable bias toward or

against acquiescense with each statement in the partition. In any case, bias toward acquiescense could only produce subadditivity, contrary bias only superadditivity, but both are found, along with additivity.

4. To ensure naturalness of formulation, we sometimes allowed nonexhaustiveness of the two alternatives in a partition. The event left out, however, was exceedingly unlikely, e.g., the extinction of the Indian rhinoceros and the Northern Pacific sea-horse at the exact same moment.
5. There is little theory or experience to guide the choice of a response or measurement method for evidence strength. A somewhat similar approach was used by Briggs and Krantz (1992). We view the method used here as far from ideal, and the later assumption of interval data (use of the numeric codes 0–5 in the table) is only partly justified. However, the satisfactory predictions shown below provide some pragmatic justification.
6. We do not know why evidence judgments seem to be more stable than probability judgments.
7. We thank the referee for these citations.

References

Baron, J. (1994). *Thinking and Deciding,* 2nd edition. New York, NY: Cambridge University Press.

Beyth-Marom, R., L. Austin, B. Fischhoff, C. Palmgren, and M. Jacobs-Quadrel. (1993). "Perceived Consequences of Risky Behaviors: Adults and Adolescents," *Developmental Psychology* 29, 549–563.

Brenner, L. and Y. Rottenstreich. (1999). "Focus, Repacking and the Judgment of Grouped Hypotheses," *Journal of Behavioral Decision Making* 12, 141–148.

Brenner, L. A. and D. J. Koehler. (1999). "Subjective Probability of Disjunctive Hypotheses: Local-Weight Models for Decomposition of Evidential Support," *Cognitive Psychology* 38, 16–47.

Briggs, L. and D. H. Krantz. (1992). "Judging the Strength of Designated Evidence," *Journal of Behavioral Decision Making* 5, 77–106.

Christensen-Szalanski, J. J. J. and C. F. Willham. (1991). "The Hindsight Bias: A Meta-Analysis," *Organizational Behavior and Human Decision Processes* 48, 147–168.

Cohen, J., E. Dearnaley, and C. Hansel. (1956). "The Addition of Subjective Probabilities: The Summation of Estimates of Success and Failure," *Acta Psychologica* 12, 371–380.

Edwards, W., H. Lindman, and L. D. Phillips. (1965). "Emerging Technologies for Making Decisions," In T. M. Newcomb (ed.), *New Directions in Psychology II,* pp. 265–325. New York: Holt, Rinehart and Winston.

Fiedler, K. and T. Armbruster. (1994). "Two Halfs May Be More Than One Whole: Category-Split Effect on Frequency Illusions," *Journal of Personality and Social Psychology* 66, 633–645.

Fischhoff, B. and W. Bruine de Bruin. (1999). "Fifty-Fifty = 50?," *Journal of Behavioral Decision Making* 12, 149–167.

Fischhoff, B., P. Slovic, and S. Lichtenstein. (1978). "Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation," *Journal of Experimental Psychology: Human Perception and Performance* 4, 330–344.

Fox, C. R. (1999). "Strength of Evidence, Judged Probability, and Choice Under Uncertainty," *Cognitive Psychology* 38, 167–189.

Kleindorfer, P., H. Kunreuther, and P. Schoemaker. (1993). *Decision Sciences: An Integrative Perspective.* New York, NY: Cambridge University Press.

Koriat, A., S. Lichtenstein, and B. Fischhoff. (1980). "Reasons for Confidence," *Journal of Experimental Psychology: Human Learning and Memory* 6, 107–118.

Macchi, L., D. Osherson, and D. H. Krantz. (1999). "Superadditive Probability Judgment," *Psychological Review* 106, 210–214.

Morgan, M. G. and M. Henrion. (1990). *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge, UK: Cambridge University Press.

Mulford, M. and R. M. Dawes. (1999). "Subadditivity in Memory for Personal Events," *Psychological Science* 10(1), 47–51.

Osherson, D. (1995). “Probability Judgment.” In E. E. Smith, and D. Osherson (eds.), *Invitation to Cognitive Science: Thinking, Second edition.* Cambridge, MA: MIT Press.

Osheson, D., D. Lane, P. Hartley, and R. Batsell. (in press). “Coherent Probability from Incoherent Judgment,” *Journal of Experimental Psychology: Applied.*, in press.

Osherson, D., E. Shafir, D. H. Krantz, and E. E. Smith. (1997). “Probability Bootstrapping: Improving Prediction by Fitting Extensional Models to Knowledgeable but Incoherent Probability Judgments,” *Organizational Behavior and Human Decision Processes* 69, 1–8.

Rottenstreich, Y. and A. Tversky. (1997). “Unpacking, Repacking, and Anchoring: Advances in Support Theory,” *Psychological Review* 104, 406–415.

Russo, J. E. and K. J. Kolzow. (1994). “Where is the Fault in Fault Trees?,” *Journal of Experimental Psychology: Human Perception and Performance* 20, 17–32.

Savage, L. J. (1954). *The Foundations of Statistics.* New York, NY: Wiley.

Tversky, A. and C. R. Fox. (1995). “Weighing Risk and Uncertainty,” *Psychological Review* 102.

Tversky, A. and D. Kahneman. (1974). “Judgment Under Uncertainty: Heuristics and Biases,” *Science* 185, 1124–1131.

Tversky, A. and D. Kahneman. (1992). “Advances in Prospect Theory: Cumulative Representation of Uncertainty,” *Journal of Risk and Uncertainty* 1, 297–323.

Tversky, A. and D. J. Koehler. (1994). “Support Theory: A Nonextensional Representation of Subjective Probability,” *Psychological Review* 101(4), 547–567.

Wallsten, T. (1971). “Subjectively Expected Utility Theory and Subjects’ Probability Estimates: Use of Measurement-Free Techniques,” *Journal of Experimental Psychology* 88, 31–40.

Wallsten, T., D. Budescu, and R. Zwick. (1992). “Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments,” *Management Science* 39, 176–190.

Windschitl, P. D. and G. L. Wells. (1996). “Measuring Phychological Uncertainty: Verbal Versus Numerical Methods,” *Journal of Experimental Psychology: Applied* 2, 343–364.

Yates, J. F. (1990). *Judgment and Decision Making.* Englewood Cliffs, NJ: Prentice-Hall.