Available online at www.sciencedirect.com
ScienceDirect
Journal of Mathematical Psychology 51 (2007) 229–241
Journal of Mathematical Psychology
www.elsevier.com/locate/jmp

Contrasting stochastic and support theory accounts of subadditivity

J. Neil Bearden[a,*], Thomas S. Wallsten[b], Craig R. Fox[c]

[a]*University of Arizona, Tucson, AZ 85721, USA*
[b]*University of Maryland, MD, USA*
[c]*University of California-Los Angeles, CA, USA*

Received 4 January 2006; received in revised form 2 January 2007
Available online 4 June 2007

Abstract

Numerous studies have found that likelihood judgment typically exhibits *subadditivity* in which judged probabilities of events are less than the sum of judged probabilities of constituent events. Whereas traditional accounts of subadditivity attribute this phenomenon to deterministic sources, this paper demonstrates both formally and empirically that subadditivity is systematically influenced by the stochastic variability of judged probabilities. First, making rather weak assumptions, we prove that regressive error (or variability) in mapping covert probability judgments to overt responses is sufficient to produce subadditive judgments. Experiments follow in which participants provided repeated probability estimates. The results support our model assumption that stochastic variability is regressive in probability estimation tasks and show the contribution of such variability to subadditivity. The theorems and the experiments focus on within-respondent variability, but most studies use between-respondent designs. Numerical simulations extend the work to contrast within- and between-respondent measures of subadditivity. Methodological implications of all the results are discussed, emphasizing the importance of taking stochastic variability into account when estimating the role of other factors (such as the availability bias) in producing subadditive judgments.

Keywords: Probability judgment; Subadditivity; Support theory

1. A stochastic model of subadditivity

People are often called on the map their subjective degree of belief to a number on the [0,1] probability interval. Researchers established early on that probability judgments depart systematically from normative principles of Bayesian updating (e.g., Phillips & Edwards, 1966). More recently investigators have observed systematic violations of the additivity principle, one of the most basic axioms of probability theory (Kolmogorov, 1933). Consider disjoint events A, B and their union $C = A \cup B$, and let Pr(.) be a normative probability measure. Additivity requires that $\Pr(A) + \Pr(B) = \Pr(C)$. For instance, if the probability that the home team wins by at least ten points, Pr(A), is .30 and the probability that the visiting team wins by at least ten points, Pr(B), is .10, then the probability that the margin of victory is at least ten points, Pr(C), must be $.40 = .30 + .10$. In contrast to this normative requirement, numerous studies have shown that judged probability, $P(.)$, is usually *subadditive* so that $P(C) < P(A) + P(B)$. Evidence of this pattern is reviewed by Tversky and Koehler (1994) and Fox and Tversky (1998). It has been observed among avid sports fans (e.g., Fox, 1999; Koehler, 1996), professional options traders (Fox, Rogers, & Tversky, 1996), doctors (Redelmeier, Liberman, Koehler, & Tversky, 1995), lawyers (Fox & Birke, 2002), and bookmakers (Ayton, 1997).

Thus far, researchers typically attribute biases in judged probability to heuristic processes based on memory retrieval, similarity judgment, or other cognitive processes (Bearden & Wallsten, 2004; Kahneman, Slovic, & Tversky, 1982; Gilovich, Griffin, & Kahneman, 2002). In contrast, a growing body of research has shown that some commonly observed patterns of bias can occur when otherwise accurate probability judgments are perturbed by random

*Corresponding author. Department of Management and Organizations, Eller College of Management, 405 McClelland Hall, Tucson, AZ 85721, USA. Fax: +520 621 4171.
E-mail address: jneilb@gmail.com (J. Neil Bearden).

0022-2496/$ - see front matter © 2007 Published by Elsevier Inc.
doi:10.1016/j.jmp.2007.04.001

error (e.g., Erev, Wallsten, & Budescu, 1994; Juslin, Olsson, & Björkman, 1997; Soll, 1996). It is important, therefore, to rule out, or at least account for, response patterns due to stochastic variability before positing complex cognitive mechanisms. The aim of this paper is to contrast explanations of subadditive judgments based on random *error* with those based on the principles of support theory (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997).[1] We provide both a formal model and two empirical studies that investigate the extent to which stochastic variability contributes to this phenomenon.

It is difficult to distinguish cognitive accounts of subadditivity from stochastic accounts of this phenomenon because past studies have asked participants to make a single probability judgment of each event A, B and $C = A \cup B$, in either between- or within-participant experimental designs. Hence, past studies have not allowed researchers to observe the variability of a particular participant's judgments of a particular event and diagnose the extent to which such variability contributes to subadditivity. In the present studies we aim to fill this gap in the literature by providing opportunities for participants to learn the relative frequencies of events in a controlled learning environment and then eliciting multiple judgments of each event.

The remainder of this paper is organized as follows. First, we summarize the major cognitive explanations of subadditivity. Next, we show how random error also might explain subadditivity, critically discuss a published test of a particular stochastic model, and develop a very general one. We then present two experiments that examine the extent to which random error can account for subadditivity, followed by a numerical simulation of our model. The paper concludes with a general discussion relating stochastic and support theory explanations of subadditivity.

Cognitive interpretations of subadditivity. Traditional accounts of subadditivity associate this phenomenon with the availability heuristic (Tversky & Kahneman, 1973). For instance, Fischhoff, Slovic, and Lichtenstein (1978) asked automobile mechanics to judge the relative frequency of different causes for a car failing to start. Mechanics estimated that a car would fail to start about 22 times in 100 due to *causes other than* the "battery, fuel system, or engine." However, when this residual category was broken down into distinct components consisting of "starting system," "ignition system," "mischief," and a new (less inclusive) residual category, a second group of mechanics reported judgments that summed to 44 times in 100. These authors argued that unpacking the catch-all category into specific instances enhanced the accessibility of particular causes and therefore their apparent likelihood (see also Russo & Kolzow, 1994; Ofir, 2000). Likewise, in support theory (Tversky & Koehler, 1994; Rottenstreich & Tversky, 1997) subadditivity is attributed to availability: unpacking a description of an event into more specific constituents may remind people of possibilities that they would have overlooked or enhance their salience.

Although availability could contribute to subadditivity in situations where a category of events is partitioned into constituents, it is unlikely to provide a satisfactory account of all instances of subadditivity. First, a number of studies have shown that when descriptions of events (e.g., "precipitation next April 1") are unpacked into a disjunction of constituents (e.g., "rain or sleet or snow or hail"), judged probability sometimes increases, but not as dramatically as the sum of the probabilities of these constituents when they are judged separately (Rottenstreich & Tversky, 1997; Fox & See, 2003). In fact, unpacking descriptions into a disjunction of constituents and a catch-all sometimes leads to a *reduction* in judged probability. For instance, Sloman, Rottenstreich, Wisniewski, Hadjichristidis, and Fox (2004) found that the median judged probability that a randomly selected death is due to "disease" was .55, the median judged probability that it is due to "heart disease, cancer, stroke, or any other disease" rose to only .60, and the median judged probability that it is due to "pneumonia, diabetes, cirrhosis, or any other disease" actually *decreased* to .40. Second, subadditivity is typically observed when a dimensional space (e.g., future daytime high temperature or closing value of a stock index) is partitioned into constituents, and it seems implausible that the availability mechanism contributes here. For example, Fox et al. (1996) found that most professional options traders in their sample judged the probability of the event "Microsoft Stock (MSFT) will close below $94 per share two weeks from today" to be lower than the sum of his or her estimated probabilities of the events "MSFT will close below $88 per share" and "MSFT will close at least $88 per share but below $94 per share."

Such observations have motivated a second cognitive interpretation of subadditivity: bias toward an "ignorance prior" probability. When respondents assign probabilities simultaneously to each of n events into which a sample space is partitioned, their responses are typically biased toward $1/n$ for each event (Fox & Clemen, 2005; cf. Van Schie & Van der Pligt, 1994). More generally, when people are asked to judge the probability of binary events, their responses are biased toward $\frac{1}{2}$, placing equal credence on the event and its complement, unless a partition of the event space into $n > 2$ interchangeable events is especially salient (Fox & Rottenstreich, 2003; see also Fischhoff & Bruine De Bruin, 1999). Thus, if probabilities of events A,

[1] Throughout we will use the terms random error, error, and stochastic variability interchangeably. In statistics, one uses the term "error" to denote the variability of realizations of random variables about their *expectation*. Here, we use the term similarly; however, we emphasize the variability of the realizations about the random variable's *median*. The reason for this alternative usage will be made clear, and formalized, below. For further clarification on the use of the term "error" in this context, see Brenner's (2000) theoretical note regarding the Erev et al.(1994) model, in which he questioned that usage; and Wallsten, Erev, and Budescu's (2000) response.

B and $A \cup B$ are all biased toward 1/2 then these probabilities will be subadditive.

Stochastic interpretations of subadditivity. Despite the appeal of availability and ignorance priors as explanations of subadditivity, it is necessary to consider the extent to which the phenomenon may be driven by simple random perturbations in judgment. Tversky and Koehler (1994) pointed out that (explicit) subadditivity can be accommodated by a stochastic model such as Erev, Wallsten, and Budescu's (1994), which hypothesizes that subjective probability estimates reflect an underlying covert value disturbed by random error. The resulting overt estimate is regressive because the probability scale is bounded at 0 and 1, thereby causing the error distribution to be skewed inward toward 0.5. Likewise, Brenner (2003) proposed a stochastic version of support theory that can accommodate subadditivity without resorting to availability- or ignorance prior-based explanations. In Brenner's model, error is attached to the process by which people recruit evidence for the target event and its complement.

For the present treatment, we investigate a simpler model of random error. We seek the most general form of the assumption that an observed probability estimate in response to an event description contains a trial-by-trial random component. Specifically, following Erev et al. (1994), we assume that $R(X)$, the estimate for a described event X, depends on a covert judgment $C(X)$ and a random component e. That is,

$$R(X) = f(C(X), e), \tag{1}$$

where f is monotonically increasing in its arguments.

In Brenner's (2003) version, the support accrued for an event description is log-normally distributed, with a consequent distribution over the probability estimate. Without taking a stand on the source of the variability, we suggest a weaker alternative to Brenner's model based on the assumption that the probability estimates, $R(X)$, have a *qualitatively symmetric error distribution*, by which we mean that the observed probability estimate is just as likely to be less than as greater than the underlying covert value. Such a distribution will have its median at the underlying covert value and be skewed inward, due to the bounds at 0 and 1; but beyond the inward skewing, it is not constrained to any particular shape or form. Given the absence of empirical evidence to the contrary, we take this to be a rather weak assumption. (It is also reasonable to assume that the distribution is single peaked. Our theorems, however, do not depend on this assumption.) The source of the stochastic variability may be the accrual of support (as Brenner assumes) or more generally the memory search that precedes a probability estimate. Alternatively, the stochastic component may be due to criterion variability in mapping the covert judgment to an overt response. (Budescu, Wallsten, & Au, 1997; Wallsten, Bender, & Li, 1999; and Wallsten & González-Vallejo, 1994, all discuss the last two causes and the difficulty in distinguishing them.) The impact of the stochastic error is the same regardless of the source.

Unlike Brenner's (2003) model we first assume additivity at the covert level. We prove that this assumption along with that of qualitatively symmetric error yields subadditivity at the overt response level when the data are summarized in terms of means. The responses are additive, however, when they are summarized by medians.

We then weaken our assumptions and allow for subadditive covert judgments. Still assuming qualitatively symmetric error, we observe a surprising result: overt subadditivity at the level of means provided the covert estimates of the subevents (and therefore the medians of the observed values) are all less than .5. The same consequence does not necessarily obtain when the covert estimate of *any* subevent is greater than .5. The medians, of course, will reflect subadditivity if the covert values do, regardless of the locations of the covert estimates.

To state the models fully and properly, it is necessary to introduce some notation and definitions. Consider an event, X, composed of mutually exclusive and collectively exhaustive subevents, X_1, X_2, …, X_n. Let $C(X_i)$ be the covert estimate and $R(X_i)$ be the overt estimate of X_i (based on the description) with $R(X_i)$ and $C(X_i)$ both bounded in the closed [0,1] interval. Let e denote the random variable representing trial-by-trial variability such that Eq. (1) holds. Finally, let $E(X_i)$ denote the mean, or the expected value, of $R(X_i)$.

A few words of explanation are in order: Because $C(X_i)$ is covert, it is not directly observable. $R(X_i)$, on the other hand, is. The mean of replicated observations of $R(X_i)$ is a sample estimate of $E(X_i)$. Under assumption A2.1 below, the median of these replicated observations is an estimate of $C(X_i)$. Finally, we explicitly do not assume additive error, only that the observed estimate is a function of the covert value perturbed by a stochastic component in some fashion.

With these definitions in hand, we make the following assumptions:

A1. *Additivity of covert judgments.* $C(X) = \sum_{i=1}^{n} C(X_i)$.

A2. *Qualitatively symmetric, skewed random error.*

A2.1. R is distributed such that $\Pr(R(X_i) < C(X_i)) = \Pr(R(X_i) > C(X_i))$.

A2.2. Define $S(X_i) = E(X_i) - C(X_i)$ as the skew of the distribution. $S(X_i)$ monotonically decreases with X_i such that $S(X_i) = 0$ if $C(\mathrm{X}_i) = .5$.

These assumptions merit some discussion. Assumption A1, that covert judgments are additive, implicitly brings with it the assumption of extensionality. (A1′ below relaxes this assumption.) According to assumption A2.1, $R(X_i)$ is equally likely to above as below $C(X_i)$. Thus $C(X_i)$ is the median of the distribution of $R(X_i)$, and as a consequence, A1 and A2.1 together imply that the median estimates should satisfy additivity. A2.1, moreover, guarantees that the (limiting) distribution of overt responses will be skewed

inward, due to the bounds at 0 and 1. Assumption A2.2 regularizes this characteristic by assuming the skew changes smoothly from positive to negative as the covert estimate increases from 0 to 1 such that it equals 0 when $C(X_i) = .5$.

Our first model differs from Brenner's (2003) in two ways. First, his does not invoke A1; more accurately, the construct of an underlying covert judgment, which is at the heart of A1, is foreign to Brenner's model. Second, his model implies a distribution over $R(X_i)$ that is inconsistent with A2.2.[2]

For our model we prove:

Theorem 1. *If A1 and A2 hold and* $C(X_i) < C(X)$ *for all* X_i, *then*

$$\sum_{i=1}^{n} E(X_i)/E(X) > 1.$$

That is, when assumptions A1 and A2 hold and in addition the estimate of each subevent is strictly less than that of the event, the means of replicated observations will exhibit subadditivity. (As already noted, and by the same assumptions, the medians are additive.) The proof of this theorem is in Appendix A.

When A1 is relaxed to allow covert subadditivity (as in Brenner, 2003, model), subadditivity of mean responses is guaranteed only when the (covert) estimate of the event is greater than 0.5 and the (covert) estimates of all subevents are less than 0.5. Specifically, Theorem 2 follows under these conditions when A1 is replaced by

$$\text{A1}'. \quad C(X) < \sum_{i=1}^{n} C(X_i).$$

Theorem 2. *If A1′ and A2 hold,* $C(X) \geqslant 0.5$ *and* $C(X_i) < 0.5$ *for all* X_i, *then,*

$$\frac{\sum_{i=1}^{n} E(X_i)}{E(X)} > \frac{\sum_{i=1}^{n} C(X_i)}{C(X)}.$$

The proof is in Appendix A. Surprisingly, when A1′ is substituted for A1 and either $C(X) < 0.5$ or $C(X_i) \geqslant 0.5$ for some i, then without stronger assumptions about the error distribution, no prediction about the additivity of the mean responses can be made. For example, they could be superadditive. The reasoning behind this statement is amplified in Appendix A.

We should note that the implications of Theorem 2 are consistent with results from simulations of Brenner's model (see, Brenner, 1995). Examining some special cases (i.e., certain parameterizations of the log-normal support distributions), he showed that average observed probability judgments were subadditive, while the median ones were additive. Theorem 2 entails that this result obtains under a broad class of (observed) response distributions.

In the section that follows we present detailed results of two experiments (and a summary of a third experiment) in which we asked participants to observe the frequency of events in a controlled learning environment and then to make multiple probability judgments of each event. The present analysis suggests the following testable hypotheses:

(1) The skew of replicated judgments changes progressively from positive to negative as the median judgment increases from 0 to 1 and is 0 when the median judgment is 0.5. (Assumption A2.2.)
(2) Interpreting response medians as sample estimates of the underlying covert estimates (Assumption 2.1), if the medians are additive, then the response means are subadditive (Theorem 1).
(3) If the response medians are subadditive, the median event estimate is greater than or equal to 0.5, and the median subevent estimates all are less than 0.5; then the means will exhibit greater subadditivity than the medians. (Theorem 2)
(4) No necessary predictions follow when the medians are subadditive and the side conditions on the median judgments do not hold. However if response variability is not too large, one would expect the same results as predicted above.

2. The experiments

Most studies of subadditivity have relied on between-respondents comparisons and even those that have used within-respondents designs failed to include replicated judgments. The experiments reported here use replicated responses on a within-respondent basis. Participants in these experiments first observed random draws from a continuous sample space in order to form an impression of its frequency distribution. They next estimated the frequency per 100 trials of an event taking on values within specified intervals. (We will refer to these judgments as probability judgment, though, technically, the participants provided relative frequency judgments.) In the response phase of Experiment 2, participants also judged whether specified event probabilities were too high or too low. We used visual stimuli distributed along a dimensional space in a controlled learning environment in order to minimize to the contribution of availability- and ignorance-prior based sources of subadditivity[3] and to

[2]We make this statement on the basis of extensive numerical calculations of the mean and skew of the $R(X_i)$ distributions given various mean-variance combinations for the lognormal distribution over support of an event and its complement, $s(X_i)$ and $s(\bar{X}_i)$, respectively. The skew of the $R(X_i)$ distribution does not monotonically decrease with the mean, but instead shows a cubic trend, with the precise shape of the function depending on the mean and variance of the lognormal distribution.

[3]We expected that unpacking events depicted as visual segments into a set of sub-segments would not have a large effect on the cumulative accessibility of instances of these events so that availability-based sources of subadditivity would be minimized. We also expected that for many

minimize the extent to which participants could recall their previous judgments of each event Thus we hoped to bring trial-by-trial variation of responses into shaper focus.

2.1. Experiment 1

2.1.1. Method

Participants and general conditions. Participants (Ps) were 41 members of the University of North Carolina-Chapel Hill community recruited by means of posters on campus bulletin boards promising payment according to performance. The experiment was computer-controlled, with Ps working individually in sound-attenuated cubicles. Stimuli were shown on 15-inch color monitors and responses were made on the keyboard.

Learning phase. A thin horizontal line spanned the computer monitor. Ps were told that the line contained two targets, called red and blue, invisible to them, and that balls aimed at the targets would appear on the line one at a time as black dots. Following each ball, they were to indicate by typing "*r*" or "*b*" at which target they thought it was aimed. Immediate feedback was provided by changing the color of the ball to red or blue. For the purpose of locating stimuli, the line was divided into 600 equal-width segments. The red and blue targets were centered within segments 272 and 367, respectively, and stimuli aimed at each were drawn from normal distributions centered at the target locations and with standard deviation of 95. The targets were thus one standard deviation apart ($d' = 1$). Participants were first presented with 150 training trials, half of which were drawn from each target distribution. The stimuli for each target were determined by dividing the line into 75 equal-probability intervals according to the operative distribution and taking the location at the center of each interval. This procedure guaranteed that all Ps would experience each target's normal distribution of ball landings. The 150 locations were presented in a random order that was shown to half of the Ps; the stimuli were presented in the opposite order to the other half of the Ps. Ps earned \$.05 for each correct response.

Estimation phase. In this phase, the Ps estimated the probabilities that a subsequent ball would fall within specific line segments. On each trial, a segment of the line was highlighted and the Ps were asked how many of the next 100 throws they expected would land within that segment. There was no time limit on responding.

Six non-overlapping, contiguous line segments (elementary events) were constructed in such a way that additivity of the responses could be assessed. For example, two contiguous elementary segments, A and B, as well as their concatenation, AB, were judged separately. The elementary elements were combined so that in total there were 14 target segments. These 14 segments (events) and their corresponding true probabilities (calculated as the mean of the relative frequencies of the segments under the red and blue target distributions) are shown in the first two columns of Table 1. Ps provided six probability estimates for each segment. To mask the repetition of the target segments, we also included 16 non-target segments, each of which was presented only one time. Thus, Ps provided a total of 100 estimates, but only 84 of them were used in the data analyses.

Table 1
Presentation Probabilities by Event and Experiment

Experiment 1		Experiment 2	
Event	Probability	Event	Probability
A	.05	A	.12
B	.13	B	.19
C	.18	C	.37
D	.37	D	.23
E	.21	E	.09
F	.06	AB	.31
AB	.17	BC	.58
CD	.54	CD	.60
EF	.27	DE	.32
ABC	.35	ABC	.68
BCD	.67	BCD	.79
CDE	.75	CDE	.69
DEF	.64	ABCD	.88
CDEF	.82	BCDE	.91

Note: Elementary events are denoted with a single letter (e.g., A); events denoted with multiple letters refer to events that are concatenations of elementary events. For example, AB is the event that consists of the concatenation of events A and B. From left to right, the position of the elementary events on the computer screen was A, B, C, D, E, F for Experiment 1 and A, B, C, D, E for Experiment 2. Indicated probabilities of concatenated events may deviate from the sum of the probabilities of the components events due to rounding errors.

2.1.2. Results

Accuracy of estimates. Although the Ps' estimates were in terms of frequencies, all results will be reported as probabilities (*estimate*/100). It is useful to obtain an overview of the data by looking at the probability estimates as a function of the objective values before considering the predictions. We calculated for each P the median and the mean of his or her 6 estimates for each of the 14 events; the open diamonds and the nearby horizontal lines, respectively, of Fig. 1 show the means over participants of these two quantities. Note first that the estimates are overall relatively accurate; the mean absolute deviations of the means and of the medians are .018 and .020, respectively. Nevertheless, there is a tendency to overestimate low and underestimate high probabilities, with the crossover at approximately 1/3. The highly accurate estimates leave little room for demonstrations of subadditivity, let alone comparisons of degrees of subadditivity. Finally, the discerning reader may notice that the medians tend to be

(*footnote continued*)
participants the most accessible ignorance priors would be the proportion of the visual sample space spanned by each event so that such ignorance priors would be additive.

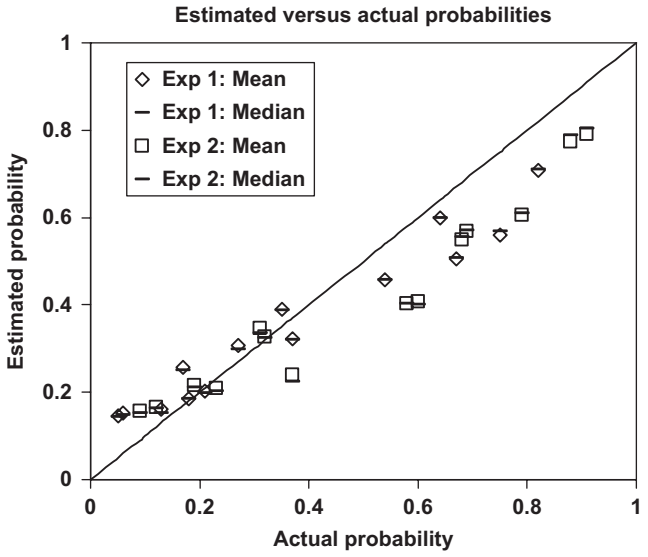

Fig. 1. Mean over Ps of the mean and median estimates, respectively, as a function of the objective event probability for Experiments 1 (open diamonds and horizontal bars) and 2 (open squares and horizontal bars).

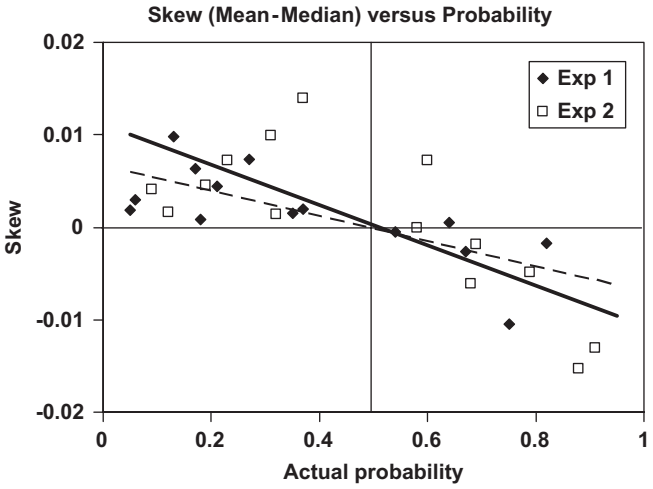

Fig. 2. Mean over Ps of the skew, defined as the mean minus the median probability estimate, as a function of the event probability, for Experiments 1 (solid diamonds and trend line) and 2 (open squares and dashed trend line).

slightly more extreme than the means, i.e., lower than the means below .5 and greater than the means above .5, consistent with assumption A2.2 regarding skew, defined for our purposes as the difference between the mean and median judgments.

Skew of response distributions. The pattern of mean–median differences evident in the group data holds for most individual Ps, as well, as established by regressing separately for each P the skew of his or her six responses for each event onto their respective true probabilities. The solid diamonds of Fig. 2 show the mean skew over Ps for each of the 14 events as a function of the event probabilities. As required by assumption A2.2, that skew decreases as the mean probability estimate increases, the linear trend is significant ($r = -.73$, $p < .001$ one-tailed), while no higher order trends are. Moreover, consistent with A2.2, the skew predicted by the linear regression line ($\hat{s} = -.014p + .007$) for $p = .50$ is .00. A2.2 in fact states that skew decreases monotonically with the estimated probability, not with an external measure such as the objective value. Standard regression analyses are not appropriate when both variables contain random error. Nevertheless and unsurprisingly given the accuracy of the estimates, the pattern does not change when skew is plotted against the mean estimates—no trend beyond the linear is significant and the best fitting linear equation predicts skew of .00 when the mean estimate is .50. Analyses of individual participant data found that the slope of the linear regression of skew onto event probabilities was negative for 27 of the 41 Ps ($p < .05$, one-tailed).

Comparison of subadditivity calculated with means and medians. Our design allowed for 15 tests of additivity for each P, as shown in column 1 of Table 3. For each P, we first calculated the log unpacking ratios (Tversky & Koehler, 1994) based on the mean and median responses to each event, i.e.,

$$\ln((mean(A) + mean(B))/mean(AB)) \quad \text{and}$$
$$\ln((median(A) + median(B))/median(AB)),$$

as well as their difference. Henceforth, we use URMN and URMD to refer to the unpacking ratios based on means and medians, respectively. Thus, we have log-URMN and log-URMD. In order to obtain more reliable estimates and to reduce the number of tests, we took the mean of the 15 log-URMN, 15 log-URMD values, and 15 differences between the two for each P. The first row of Table 3 shows the group means and standard deviations of these statistics. The mean log-URMD indicates significant subadditivity ($t(40) = 5.05$, $p < .001$). Because the constituent probabilities (with a few exceptions) are less than 0.5, Theorem 2 is operative for prediction purposes. Accordingly, repeated-measures t-test reveal that the log-URMN for each event are significantly greater than the log-URMD ($t(40) = 1.86$, $p = .035$ one-tailed). Not surprisingly, the mean of the log-URMN indicates significant subadditivity ($t(40) = 5.34$, $p < .001$).

2.1.3. Discussion

Despite the overall accuracy of the mean and median estimates, and the remarkably little skew in the distributions of repeated responses to an event (from roughly 0.01 to roughly 0.02, cf. Fig. 2), the data confirm our predictions. Skew monotonically decreased with median estimate and was 0 at median estimate of 0.5, consistent with A2.2. Because medians displayed subadditivity and side conditions were essentially met, Theorem 2 applied. As predicted, subadditivity of means was greater than subadditivity of medians.

It is important to emphasize that the degree of subadditivity overall was quite small in this instance. Converting the mean values from logs, the mean unpacking ratio for the mean estimates was 1.15 and that for the medians was 1.14. We suspected that Ps exhibited a high degree of accuracy in Experiment 1 because of the extensive training in which they received immediate feedback and payoffs for correct predictions. We therefore ran a follow-up study in which 112 participants observed graphic events in rapid succession but did not make predictions or receive feedback during the learning phase. The 2×2 between-participant design manipulated number of training trials (100 versus 300) and presence versus absence of response-contingent (incentive-compatible) payoffs for the subsequent estimates. Surprisingly, there were no significant effects of the between-respondent variables. In all respects, the results mirrored those of the first study and therefore we omit details of this study.

2.2. Experiment 2

Our second experiment was designed to probe the covert-overt response distinction. Participants completed a training phase followed by two data-collection tasks. First, we asked Ps to provide probability estimates, as in Experiment 1. Second, we asked Ps to indicate whether various probabilities were greater than or less than the true probability for the interval being displayed (unbeknownst to Ps the probabilities presented were their own median estimate, mean estimate, and the objective value). We assume that such binary responses are less biased than judged frequencies, since they do not require mapping the covert judgment onto a verbal (or numerical) overt response (see Erev et al., 1994, for a discussion of the sources of response variability). Because our model predicts that the mean of a sample of judgments should be more regressive (i.e., closer to 0.50) than the median, we expected that the proportion of median judgments labeled as "too high" should be closer to 50% than the proportion of mean judgments.

2.2.1. Method

Participants. We recruited 61 undergraduate students from introductory psychology courses at UNC-CH, who received credit toward a course requirement for their participation.

Learning phase. Participants were provided a cover story that described a farmer's attempt to catch a rat that had been a nuisance. The farmer wanted to better understand the rat's behavior in order to catch him. Ps were told that the farmer had enlisted them to observe a field to learn where and how often the rat appeared in different areas. They were to carefully observe that rat's behavior and to learn where and how often it appeared. Ps learned the stimulus distribution by observing an animated rat that randomly popped up and down in a one-dimensional field, alternating between 1 s appearances and disappearances. We created the distribution of appearances by dividing the field into five non-overlapping, contiguous segments of equal length. Within each segment the distribution of appearances was uniform, but over the entire field it was roughly symmetric and single-peaked. The true probabilities (i.e., relative frequencies) for each elementary event and for the concatenations of the elementary events that were judged are shown in the third and fourth columns of Table 1. The order of appearance was randomized for each P. All Ps experienced 100 learning trials, with the order of presentation individually randomized. No payoffs were used.

Frequency estimates. After learning the stimulus distribution, Ps read a cover story describing the farmer's plan for catching the rat. They were told that he wanted to build a fence somewhere in the field to trap the rat; and that their task was to estimate the number of times that the rat would appear within the fence, which was to span two horizontally separated fence poles, in its next 100 appearances.

We created different fence positions by sectioning the field into regions based on the 5 non-overlapping, contiguous equal-length (elementary) segments used to construct the presentation distribution. Ps reported estimates for both the elementary segments and for all possible concatenations of adjacent segments.

Ps estimated the frequencies of each of the 14 events 7 times. The stimuli were presented in 7 blocks under the constraints that each position was presented only once within each block and no position was judged consecutively. With the addition of 22 non-replicated trials, each P provided a total of 120 estimates. No time limit was imposed on the responses.

Choice responses. After the frequency estimation task, Ps were asked to evaluate the responses of "another observer." They were shown each of the 14 fence positions that they had repeatedly judged in the frequency estimation task along with an estimate of the number of times that the rat would appear within the fence in its next 100 appearances. Their task was to respond whether the estimate was too high or too low. In actuality, the responses shown to each P were his/her own mean or median responses to each event, or the actual probabilities of each event. Thus Ps were presented each position 3 times, for a total of 42 choices. The presentation order was randomized for each P.

2.2.2. Results

Accuracy of estimates. As before, we present all frequency/100 estimates as probabilities. For each P, we calculated median and the mean estimate for each of the 14 events. The open squares and the nearby horizontal lines, respectively, of Fig. 1 show the means over Ps of the mean and median estimates. The results mirror those of Experiment 1 almost perfectly: Estimates are very accurate, slightly overestimate the low probabilities, cross the diagonal at around 1/3, and demonstrate a slight inward skew (means slightly more regressive than the medians).

Skew of response distributions. The open squares of Fig. 2 show the mean skew (mean-median) over Ps as a function of the objective probabilities for Experiment 2. The linear correlation is .74 ($p<.001$) and is negative for 48 of the 61 Ps ($p<.0001$, one-tail). The quadratic component is significant ($p<.03$), decreasing the multiple correlation to −.90, but the cubic trend is not. The patterns are unchanged when mean skew is plotted as a function of the mean probability estimates instead of the objective values.

Comparison of subadditivity calculated with means and medians. The current design allowed 27 tests of additivity for each P, as summarized in columns 2 and 3 of Table 2. For each P, we calculated the 27 values of log-URMN, log-URMD, and of their differences. We then took the within-P means of these three statistics. Row 2 of Table 3 shows the group means and standard deviations of these within-P means. The mean of the log-URMD is not significantly different from 0 ($t(60) = -0.57$, *ns*), suggesting additivity and bringing Theorem 1 into play. On that basis, we expect mean estimates to be subadditive. In fact, repeated-measures t-tests revealed that the mean log-URMN is significantly greater than mean log-URMD ($t(60) = 4.99$, $p<.001$).

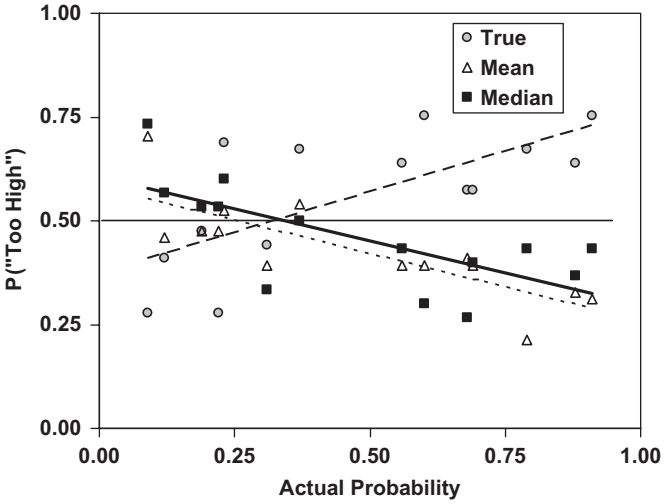

Fig. 3. Mean over Ps of the proportion of "too high" responses when the participants' mean (open triangles and dotted trend line) or median (solid squares and trend line) estimate or the actual probability (gray circles and dashed trend line), respectively, was shown, as a function of the event probability, for Experiment 2.

Table 2
Additivity checks by experiment

Experiment 1	Experiment 2	
(A + B)/AB	(A + B)/AB	(A + B + C + D)/ABCD
(C + D)/CD	(B + C)/BC	(AB + C + D)/ABCD
(E + F)/EF	(C + D)/CD	(ABC + D)/ABCD
(A + B + C)/ABC	(D + E)/DE	(AB + CD)/ABCD
(AB + C)/ABC	(A + B + C)/ABC	(A + BC + D)/ABCD
(B + C + D)/BCD	(AB + C)/ABC	(A + BCD)/ABCD
(B + CD)/BCE	(A + BC)/ABC	(A + B + CD)/ABCD
(C + D + E)/CDE	(B + C + D)/BCD	(B + C + D + E)/BCDE
(CD + E)/CDE	(BC + D)/BCD	(BC + D + E)/BCDE
(D + E + F)/DEF	(B + CD)/BCD	(BCD + E)/BCDE
(D + EF)/DEF	(C + D + E)/CDE	(BC + DE)/BCDE
(C + D + E + F)/CDEF	(CD + E)/CDE	(B + CD + E)/BCDE
(C + DEF)/CDEF	(C + DE)/CDE	(B + CDE)/BCDE
(CDE + F)/CDEF		(B + C + DE)/BCDE
(CD + EF)/CDEF		

Table 3
Group means (and standard deviations) of individual mean and median log unpacking ratios and of their difference

Experiment	Number of Ps	Number of tests/P	Log unpacking ratio based on		Difference
			Means	Medians	
1	41	15	.140** (.192)	.129** (.192)	.011* (.038)
2	61	25	.020 (.172)	−.013 (.169)	.033** (.052)

Note: Each P contributed one observation to each mean and standard deviation. This observation was the mean of the corresponding statistic over the number of tests per P.
*$p<.05$, one-tailed.
**$p<.001$, one-tailed.

Choice responses. Assuming that Ps overestimate low and underestimate high probabilities but make relatively unbiased choices comparing estimates to true probabilities, we predicted that the proportion of times mean and median judged probabilities are judged to be "too high" will be above 50% for low probabilities and below 50% for high probabilities (see Fig. 1). This was indeed confirmed in the data as seen in Fig. 3. Moreover, if means are more regressive than medians, as assumed, then the proportions of "too high" choices should be closer to 50% for medians than for the means. Again, Fig. 3 shows this pattern. The negative slope for the choices regarding means ($b = -0.32$, $r = -0.80$; $p<0.001$) is consistent with the assumed regressive nature of the mean estimates. However, we did not expect the same slope for choices regarding medians ($b = -0.31$, $r = -0.70$, $p<0.01$). Finally, turning to choices concerning objective probabilities, we find that Ps generally judged low probability events to be "too low" and high probability events to be "too high" ($b = 0.39$, $r = 0.70$, $p<0.01$). Thus there seems to be evidence of regressive bias in judged probabilities even when a judgment on the [0,1] interval is not explicitly called for. Interestingly, the crossover from below to above 50% is roughly at 0.3 rather than at 0.5.

2.2.3. Discussion

Unlike Study 1 in which median estimates were significantly subadditive, median estimates in Study 2 did not significantly differ from additivity. However, like Study 1, mean estimates in Study 2 were significantly more subadditive than median estimates. Thus, despite the relative

coherence of estimates in Study 2, the data again support the role of regressive error in the estimation process.

The choice results are particularly interesting. They demonstrate in a manner uncontaminated by any mapping to an explicit [0,1] response interval an overestimation of low and an underestimation of high probability events. The crossover in the neighborhood of .3 is surprising.[4] Moreover, choice results show that mean estimates are on average regressive relative to medians, which under our model are estimates of corresponding covert judgments. Surprisingly, median judged probabilities are also regressive. One possible explanation is that assumption A2.1, qualitatively symmetric random error, fails in favor of an error model that provides greater than 50% of the error in the regressive direction. Alternatively, it may be that covert judgments differ as a function of the response mode (i.e., probability assessment versus choice). Either way, the data seem to support the notion that some instances of subadditivity can be partly attributed to stochastic variability.

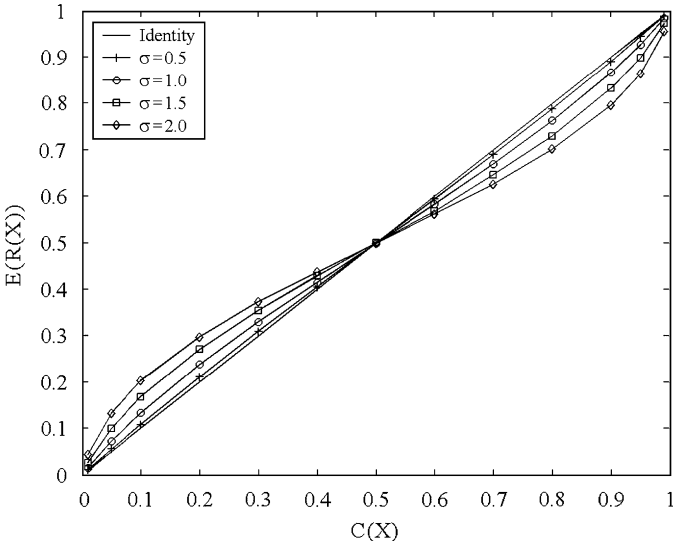

Fig. 4. Expected overt probability judgment as a function of covert judgment for various degrees of random error.

2.3. Numerical extensions of theoretical results

Our theoretical and empirical results regarding additivity are based on intra-individual (i.e., within-respondent) subadditivity. Most previous research, however, has tested additivity using between-respondents methods. In this section, we report results from numerical simulations that allow us to compare within- and between-respondent measures of additivity based on mean and median judgments.

Recall that under our second assumption (A2), response error is qualitatively symmetric: overt judgments are as likely to be greater than as less than their respective covert judgments. In order to derive numerical predictions regarding within- and between-respondents effects of error, however, we must further specify the nature of the random error. Following, Erev et al. (1994) as well as Brenner (2003), let us suppose that the random error is normally distributed with mean 0 and standard deviation σ in log-odds space. Specifically, let

$$R'(X) = \ln\left(\frac{C(X)}{1-C(X)}\right) + e, \tag{2}$$

where e is i.i.d. with $E(e) = 0$ and $Var(e) = \sigma^2$. Then, returning to probability space, we get

$$R(X) = \frac{\exp(R'(X))}{1+\exp(R'(X))}. \tag{3}$$

When $\sigma > 0$, $E(R(X)) > C(X)$ if $C(X) < .5$; $E(R(X)) = C(X)$ if $C(X) = .5$; and $E(R(X)) < C(X)$ if $C(X) > .5$. However, because the error is symmetric (with mean 0), the median $R(X) = C(X)$ for all $C(X)$. $E(R(X))$ as a function of $C(X)$ is shown in Fig. 4 for several values of σ. Note that the curves become more regressive as σ increases. We invoked Eqs. (2) and (3) to compare within- and between-respondent effects of error using the procedures described next.

Generating simulated judgments. For each simulated decision maker (DM), we randomly and independently sampled k values from a uniform distribution on the interval $(0, C(X))$. These values, multiplied by a constant α, correspond to the covert judgments, $C(X_i)$, for k subevents that together form the event X, subject to the constraint that

$$\sum_{i=1}^{k} C(X_i) = \beta C(X).$$

Varying β allows us to make the covert judgments additive ($\beta = 1$), subadditive ($1 < \beta$), or superadditive ($0 < \beta < 1$), since

$$\frac{\sum_{i=1}^{k} C(X_i)}{C(X)} = \beta.$$

The constant α provides the means to achieve the desired value of β. Specifically, let y_i denote the values sampled on $(0, C(X))$. Then $\alpha = \beta C(X)/\sum_{i=1}^{k} y_i$ and $C(X_i) = \alpha y_i$.

Under this scheme, we can manipulate the value of the covert judgment for the global event, X, by tuning $C(X)$; and all sets of feasible $C(X_i)$ are sampled with equal probability. Then, to obtain the simulated overt judgments we transform the $C(X_i)$ and $C(X)$ according to Eqs. (2) and (3), using a common σ, to get the simulated overt judgments $R(X)$ and $R(X_i)$ ($i = 1, \ldots, k$). Finally, the additivity of the simulated judgments is assessed with the unpacking ratio

$$UR(C(X), \beta, \sigma) = \frac{\sum_{i=1}^{k} f_m(\{R(X_i)\})}{f_m(\{R(X)\})},$$

[4] It is possible that the crossover point near .3 in Study 2 could reflect a bias toward an ignorance prior probability of 1/3 when the field is partitioned into three segments by the placement of two fence posts (cf. Fox & Rottenstreich, 2003; Fox & Clemen, 2005; See et al., 2006).

where $f_m(.)$ returns the mean of its argument if $m = 0$ and returns the median if $m = 1$. Within-respondent additivity can be measured by sampling multiple overt judgments, the $R(X)$ and $R(X_i)$, for each simulated DM and then taking means and medians of these judgments to compute the *UR*. Sampling the overt judgments from multiple (simulated) DMs whose $C(X_i)$ may vary but who share a common $C(X)$ allows us to examine the between-respondents case.[5]

Simulation design. For both the within- and between-respondents tests we factorially combined the number of evaluated subevents $k = \{2,4,8\}$, the value of the covert judgment $C(X) \in \{.10, .20, \ldots, .90\}$, the additivity of the covert judgments $\beta \in \{.75, .95, 1.00, 1.05, 1.25\}$, and the degree of random response error $\sigma \in \{.50, 1.00, 1.50\}$. For the *between-respondent tests*, we independently generated one set of covert judgments for the sub-events $C(X_i)$ and their corresponding overt judgments $R(X_i)$ for each simulated DM in each cell. That procedure yielded a distribution of $R(X_i)$ for each i in which each DM contributed one observation. Taking means and medians of the resulting distributions, we calculated the URs. This procedure captured the pooling method used in most published work in which subadditivity has been assessed. For the *within-respondent tests*, we generated one set of $C(X_i)$ for each simulated DM. Then, holding the $C(X_i)$ fixed for the DM, we generated a distribution of $R(X_i)$ and computed the resulting log-URMN and log-URMD values. We repeated this process in each cell for a large number of DMs and averaged the results.

2.4. Results

Within-respondents. The within-respondent results show that subadditivity on URMN values increases as the number of evaluated sub-events k increases, and as the UR of the underlying covert judgments increases (as governed by β). Most important, we observe that the degree of random response error σ strongly affects the subadditivity of responses. Specifically, as the error in responses increases and individual responses become more regressive, subadditivity increases. Interestingly, as shown in Fig. 5 for $\sigma = 2$, we also observe that URMN is not monotonic in the covert judgment, $C(X)$, of the event X. Recall that the assumptions upon which our theorems are based did not allow us to derive predictions for cases in which not all $C(X_i) < 0.5$. In the numerical simulations, we did not constrain the $C(X_i)$ to be below 0.5, and, in fact, for the cases in which $C(X) > 0.5$ some of the simulated $C(X_i)$ met or exceeded 0.5. When this occurred, the effects of error on the $R(X_i)$ corresponding to $C(X_i) < 0.5$ was counteracted in the UR: The regressive property of the error in part "cancelled out," which accounts for decrease in the UR around $C(X) = 0.5$ in URMN seen in Fig. 5.

[5]More complicated procedures can be used to examine the more general case where both the $C(X)$ and $C(X_i)$ are free to vary across DMs by, for example, specifying a distribution from which the $C(X)$ are sampled; however, the number of additional assumptions required to do so (e.g., the nature of the $C(X)$ distribution), we believe, makes the results of such an exercise perhaps less general.

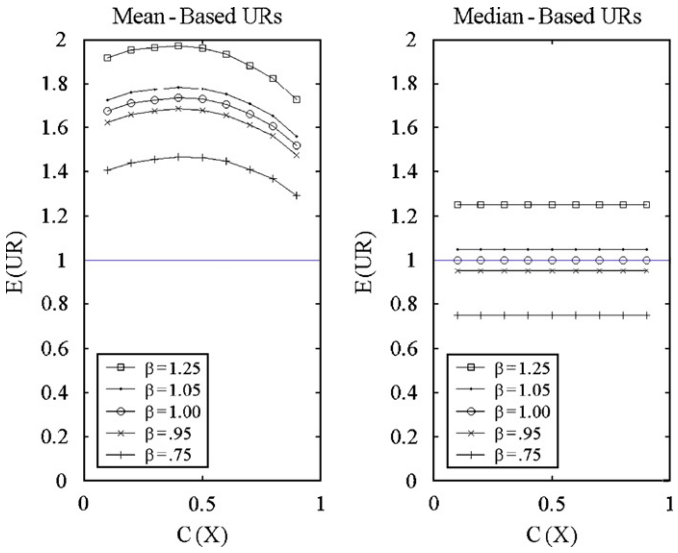

Fig. 5. Expected unpacking ratio as a function of underlying probability judgment of event X and additivity of underlying judgments (β) computed using mean (URMN—left panel) and median (URMD—right panel) within-respondent judgments for $k = 4$ and $\sigma = 2$.

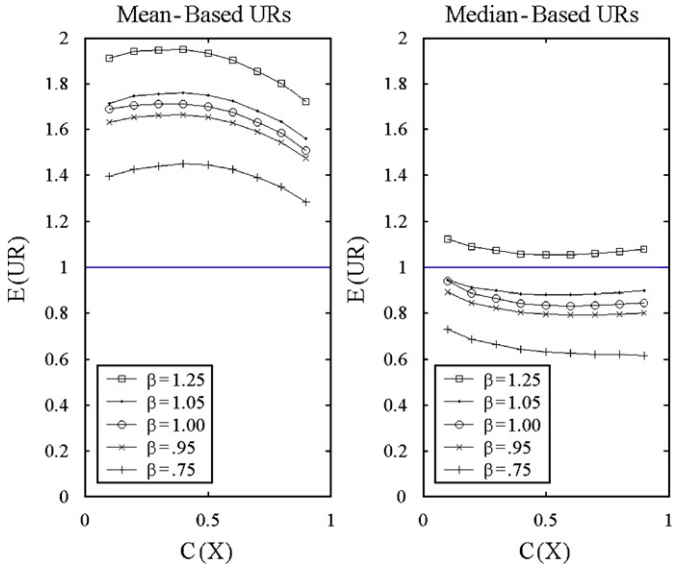

Fig. 6. Expected unpacking ratio as a function of underlying probability judgment of event X and additivity of underlying judgments (β) computed using mean (URMN—left panel) and median (URMD—right panel) between-respondent judgments for $k = 4$ and $\sigma = 2$.

The between-respondent results based on medians are affected only by the additivity of the covert judgments β. This result obtains because the error is qualitatively symmetric and therefore the median $R(X_i)$ equals $C(X_i)$ and thus UR $= \beta$.

Between-respondents. The patterns of between-respondent results on mean-based URs, shown in Fig. 6, are identical to those for the within-respondent case, though

the URs are slightly smaller in the former case. The results from median-based URs are very different from those shown in Fig. 5, however. In all cases, the median-based URs are smaller than the mean-based URs. Most interestingly, the median-based URs for the between-respondent case are smaller than β, the UR of the underlying covert judgments. It is also noteworthy that the observed UR demonstrates superadditivity for most values of β and is subadditive only when $\beta = 1.25$ (for the parameters studied here). Unlike the mean-based URs, which are strongly affected by random error, the median-based URs do not shift with random error and increase at a rather slow rate as the number of evaluated subevents increases. Overall, medians seem to provide a more robust measure of the additivity of the underlying judgments.

2.5. Discussion

The results of our simulation suggest that the summary statistics used to report judgment results can systematically affect conclusions about the degree to which the judgments show subadditivity: When means are used subadditivity is likely to appear more pronounced than when medians are used. This conclusion holds for statistics based on both within- and between-participant responses. In the original support theory article, Tversky and Koehler (1994) primarily based their conclusions regarding the subadditivity of probability judgments on mean between-respondent responses (see, for example, their Tables 1, 3–7). Rottenstreich and Tversky (1997), in their extension of the original support theory, exclusively reported median between-respondent results (see their Tables 1 and 2, and 4 and 5). Both the original paper and the extension report consistent findings of subadditivity. A survey of subsequent research on support theory finds that (between-respondent) means are reported by a number of researchers (e.g., Brenner & Rottenstreich, 1999; Koehler, Brenner, & Tversky, 1997; Koehler, White, & Grondin, 2003), though several report medians between-respondent (e.g., Fox & Birke, 2002) or compute median subadditivity within-respondent (e.g., Fox, 1999; Tversky & Fox, 1995). Because several of these studies relied on between-participant comparisons, their estimates of additivity may be biased. In particular, using means may have overestimated subadditivity (when it was found), while those reporting medians may have underestimated subadditivity.

3. General discussion

It is necessary to emphasize that in the present studies we relied on well-learned probability distributions for which we elicited multiple judgments of each event on a within-participant basis. Such a design was mandated by the questions we were investigating, because (1) it allowed us to obtain replicated observations from each individual, (2) the visual depiction of target events made it unlikely that participants would recall their prior responses and therefore be consistent for uninteresting reasons, and (3) there was a "true" probability for every event. However, such a design probably weakened our results because (1) the repeatable events used in these experiments may have given rise to less trial-by-trial variability than is commonly obtained with paradigms that use more naturalistic events (for which the construction of beliefs is a more error prone process) or elicit a single judgment for each event (for which the memory of previous responses to the same question are not available), and (2) participants may have learned the true distributions so well that they were less susceptible to response variability than is typically present in studies of subadditivity. Thus, the very design that allows us to investigate our hypotheses about the effects of random noise on additivity may also serve to reduce its magnitude. Nevertheless, the pattern of results that we obtained conforms closely with the predictions that follow from our model, supporting the notion that stochastic variability contributes significantly to observed subadditivity. The numerical simulation amplifies this result using error variances that we suspect will be more typical of judgment in general knowledge or memory retrieval contexts.

Summary of results. Although participants were relatively accurate in assessing probabilities in both experiments, they did consistently overestimate low and underestimate high probabilities, as would be expected given an inwardly-skewed error distribution. Surprisingly, the function crossed the diagonal at about 0.3 rather than at 0.5 in both cases. Also, as predicted from the assumption of an inwardly-skewed error distribution, skew defined within assumption A2.2 was positive for low probabilities, decreased monotonically to 0 at probability 0.5, and continued monotonically decreasing to negative values for high probabilities.

The probability-estimate data supported our predictions with regard to the relationship between means and medians. Specifically, mean estimates were consistently more regressive, and they also displayed more subadditivity than the median estimates. The within-respondent simulations yielded precisely the same pattern of results and also illustrated the consequences predicted under Theorem 2. When applied to between-respondent designs, and maintaining our model assumptions, the simulations demonstrate that summaries based on means are likely to overestimate subadditivity and those based on medians are likely to underestimate subadditivity, though the bias in the latter case is of a smaller magnitude.

The choice data in Experiment 2 turned out largely as predicted, with the interesting anomaly that the choice functions all crossed the 50% point, i.e., changed from a majority to a minority of "too high" choices, at a probability of about 0.3 rather than 0.5. Specifically, when shown the true probabilities, participants judged low values as too low up to a probability of about 0.3 and then judged them as too high. Conversely, when shown their mean or median estimates, they judged low values as too high up to

a probability of about 0.3 and then judged them as too low. The proportion of "too high" choices is closer to 50% for the medians than for the means, as we had expected, but the function is not flat at 50%, as predicted by the assumption of qualitative symmetric error. The pattern is consistent, however, with our overall stochastic framework, replacing the assumption of qualitative symmetric error with one that assumes heavier inward tails.

Stochastic and cognitive causes of subadditivity: *Complementary mechanisms.* The simple conclusion from our results is that overt probability estimates show trial-by-trial variability such that the distribution of estimates is skewed inward. We have shown that, in many cases, this is sufficient to produce observable probability estimates that are subadditive, even when the underlying judgments are additive. To be clear: We are not claiming that underlying judgments *are* in fact additive; rather, we simply wish to point out that response variability can contribute to observed subadditivity. It is possible to have subadditive judgments that are driven by availability—i.e., by probability judgments assigned to explicitly described events exceeding those assigned to less explicit, more difficult to think about events—and also by stochastic response variability. The same is true for subadditive judgments driven by a bias toward an "ignorance prior" probability (Fox & Rottenstreich, 2003; Fox & Clemen, 2005; See et al., 2006); these may also be made *more subadditive* by random response variability.

Our results show clearly that probability judgments tend to be regressive. However, our model is agnostic on the source of the variability and the reason for the regressive skew. Perhaps this variability and its characteristics can be captured by a computational (process) model of judgment such as MINERVA-DM (Dougherty, Gettys, & Ogden, 1999). Bearden and Wallsten (2004) showed that MINERVA-DM can provide a good model of the support accrual process in probability judgment, but they did not explore the issue of response variability. We hope that the theoretical and empirical results reported in the current paper will encourage additional work on process accounts of probability judgment.

Methodological considerations. The numerical results show that the unpacking ratio (UR)—the index of subadditivity we have used throughout is less biased in both within- and between-respondents tests when it is computed based on median judgments. We found that the expected value of URMD equals the true UR in within-respondents tests (assuming qualitatively symmetric random response error). Moreover, in the between-respondents tests, the URMD was, on average, closer to the true UR than was URMN. Further, the URMD was subadditive when and only when the true underlying judgments were subadditive in within-respondents tests, and only when the judgments were *strongly* subadditive in the between-respondents tests. Interestingly, the between-respondent URMD was biased downward: It tended to underestimate subadditivity, whereas the URMN tends to overestimate subadditivity. Researchers who wish to draw conclusions about the actual degree of subadditivity in the *judgment processes* of their participants should consider these properties of the unpacking ratio.

In sum, we argue that the way in which judgment data are analyzed affects the conclusions that are reached and have presented results showing that the conclusions are likely to differ depending on the summary statistic (mean or median) used to assess subadditivity. Our results show that the median, though still biased in between-respondents comparisons, is more robust to random response error. Of course, the statistic one uses to assess subadditivity should depend on one's research question.

Acknowledgment

We thank George Wu for stimulating and useful discussions throughout the early stages of this research. We would also like to thank Lyle Brenner for his valuable suggestions.

Appendix A

Proof of Theorem 1. $C(X) = \sum_{i=1}^{n} C(X_i)$ from A1. Substituting terms from the definition of skew leads to $E(X) - S(X) = \sum_{i=1}^{n} E(X_i) - \sum_{i=1}^{n} S(X_i)$, which is rearranged to $\sum_{i=1}^{n} E(X_i) = E(X) + \sum_{i=1}^{n} S(X_i) - S(X)$.

Note from A2.2 that $S(X_i) < S(X)$, because $C(X_i) < C(X)$. Therefore, $\sum_{i=1}^{n} S(X_i) - S(X) > 0$.

Thus $\sum_{i=1}^{n} E(X_i) > E(X)$ and therefore $\sum_{i=1}^{n} E(X_i)/E(X) > 1$.

Proof of Theorem 2. From the definition of skew,

$$\sum_{i=1}^{n} \frac{E(X_i)}{E(X)} = \sum_{i=1}^{n} C(X_i) + \sum_{i=1}^{n} \frac{S(X_i)}{C(X) + S(X)}.$$

To prove that

$$\sum_{i=1}^{n} C(X_i) + \sum_{i=1}^{n} \frac{S(X_i)}{C(X) + S(X)} > \sum_{i=1}^{n} \frac{C(X_i)}{C(X)}$$

as required, multiply the denominators out, subtract the common term from both sides of the inequality. The result is

$$C(X) \sum_{i=1}^{n} S(X_i) > S(X) \sum_{i=1}^{n} C(X_i). \tag{A1}$$

Because $C(X) \geqslant .5$, A2 guarantees that $S(X) \leqslant 0$ Similarly, because $C(X_i) < .5$ for all X_i, all $S(X_i) < 0$ Thus, expression (A1) is always satisfied.

Predictions when covert judgment is non-additive, but do not conform to the restrictions of Theorem 2. Here we show that additivity predictions are not possible unless $C(X) \geqslant .5$ and $C(X_i) < .5$ for all X_i. Consider first the case where $C(X) \geqslant .5$ and $C(X_i) \geqslant .5$ for a single X_i. Now, because $\sum_{i=1}^{n} C(X_i) > C(X)$ by A1′, the $\sum_{i=1}^{n} S(X_i)$ must be

sufficiently large to counter-balance the negative $S(X)$. But a single $S(X_i)$ is itself negative and there is no guarantee the sum of the remaining $S(X_i)$ will be capable of making up the difference.[6]

Similarly, when $C(X)<.5$, its skew is positive, but it cannot be guaranteed that the sum of the skews of the subevents will be sufficiently greater than $S(X)$ that expression (A1) holds.

References

Ayton, P. (1997). How to be incoherent and seductive: Bookmakers' odds and support theory. *Organization Behavior and Human Decision Processes*, *72*, 99–115.

Bearden, J. N., & Wallsten, T. S. (2004). MINERVA-DM and subadditive frequency judgments. *Journal of Behavioral Decision Making*, *17*, 349–363.

Brenner, L. A. (1995). A stochastic model of the calibration of subjective probabilities. Unpublished doctoral dissertation. Stanford University.

Brenner, L. A. (2000). Should observed overconfidence be dismissed as a statistical artifact? Critique of Erev, Wallsten, and Budescu (1994). *Psychological Review*, *107*, 943–946.

Brenner, L. A. (2003). A random support model of the calibration of subjective probabilities. *Organizational Behavior & Human Decision Processes*, *90*, 87–100.

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

Budescu, D. V., Wallsten, T. S., & Au, W. (1997). On the importance of random error in the study of probability judgment. Part II: Using the stochastic judgment model to detect systematic trends. *Journal of Behavioral Decision Making*, *10*, 173–188.

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review*, *101*, 519–527.

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty-fifty = 50%? *Journal of Behavioral Decision Making*, *12*, 149–163.

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.

Fox, C. R. (1999). Strength of evidence, judged probability, and choice under uncertainty. *Cognitive Psychology*, *38*, 167–189.

Fox, C. R., & Birke, R. (2002). Forecasting trial outcomes: Lawyers assign higher probability to possibilities that are described in greater detail. *Law and Human Behavior*, *26*, 159–173.

Fox, C. R., & Clemen, R. T. (2005). Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*, *51*, 1417–1432.

Fox, C. R., Rogers, B. A., & Tversky, A. (1996). Options traders exhibit subadditive decision weights. *Journal of Risk and Uncertainty*, *13*, 5–17.

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*, 195–200.

Fox, C. R., & See, K. E. (2003). Belief and preference in decision under uncertainty. In D. Hardman, & L. Macchi (Eds.), *Thinking: Psychological Perspectives on Reasoning, Judgment, and Decision Making*. Hoboken, NJ: Wiley.

[6]Even adding the assumption of symmetric skew around .5 is not sufficient to prove the result, because there is no constraint on the $C(X_i)$ and therefore no way to assure a lower bound on the sum of the skews.

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

Gilovich, T, Griffin, D., & Kahneman, D., (Eds.) (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.

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

Kahneman, D., Slovic, P., & Tversky, A. (Eds.), (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

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

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

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

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.

Ofir, C. (2000). Ease of recall vs recalled evidence in judgment: Experts vs laymen. *Organizational Behavior & Human Decision Processes*, *81*, 28–42.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

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

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

Russo, J. E., & Kolzow, K. J. (1994). Where is the fault in fault trees? *Journal of Experimental Psychology: Human Perception & Performance*, *20*, 17–32.

See, K. E., Fox, C. R., & Rottenstreich, Y. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *32*, 1385–1402.

Sloman, S., Rottenstreich, Y., Wisniewski, E., Hadjichristidis, C., & Fox, C. R. (2004). Typical versus atypical unpacking and superadditive probability judgment. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *30*, 573–582.

Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior & Human Decision Processes*, *65*, 117–137.

Tversky, A., & Fox, C. R. (1995). Weighing risk and uncertainty. *Psychological Review*, *102*, 269–283.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232.

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

Van Schie, E. C., & Van der Pligt, J. (1994). Getting an anchor on availability in causal judgment. *Organizational Behavior & Human Decision Processes*, *57*, 140–154.

Wallsten, T. S., Bender, R. H., & Li, Y. (1999). Dissociating judgment from response processes in statement verification: The effects of experience on each component. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 96–115.

Wallsten, T. S., Erev, I., & Budescu, D. V. (2000). The importance of theory: Response to Brenner (2000). *Psychological Review*, *107*, 947–949.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review*, *107*, 490–504.