Interevent relationships and judgment under uncertainty: Structure determines strategy

ALAN G. SANFEY
Princeton University, Princeton, New Jersey

and

REID HASTIE
University of Chicago, Chicago, Illinois

A fundamental empirical question regarding judgments about events is whether experienced absolute frequencies or relative frequencies are relied on when the likelihood of a particular occurrence is judged. The present research explicates the conditions under which people rely on remembered raw absolute frequencies versus on inferred relative frequencies or proportions when making predictions. Participants saw opinion poll results for candidates prior to an election and, on the basis of these, made judgments concerning the likelihood of each candidate's winning this election. Certain candidates demonstrated a high absolute frequency of winning in the polls, whereas other candidates had high relative win frequencies. The results indicated that adults are cognitively flexible with regard to the inputs used in this judgment. Certain stimulus event configurations induced reasoning by way of absolute frequencies, whereas other configurations elicited judgments based on relative frequencies. More specifically, as the relational complexity of the event structure increased and more inferences were required to make predictions, the tendency to rely on absolute, as opposed to relative, frequencies also increased.

Many common everyday judgments are concerned with the occurrence of uncertain events. Will it rain tomorrow? Which team is more likely to win the game? Will the stock price rise or fall? Often judgments of these events are expressed in the form of probability estimates: Weather forecasters give us the probability of rain, bookies give us the odds of the home team's beating the visiting team, market analysts estimate the direction of stocks. These kinds of statements reflect an underlying strength of belief about the likelihood of occurrence of the events. Cognitive scientists and, before them, philosophers have developed theoretical frameworks to describe the sources of elementary beliefs and the manner in which they are inferentially combined.

Most theorists assume that the strengths of beliefs concerning occurrences of elementary events are primitives in the analysis and then focus on describing the process of integrating the strengths. There is a general consensus that the integration of subjective probabilities, or strengths of beliefs, follows a ratio rule, usually attributed to Luce's *choice axiom* (Luce, 1959; similar formulations have been proposed in many algebraic approaches—e.g., information integration theory, Anderson, 1981, p. 77ff, and Leon & Anderson, 1974; fuzzy prop theory, Bower & Heit, 1992, and Oden & Lopes, 1997). Within the field of judgment and decision making, support theory is the best-known example of this ratio-of-primitive-strengths approach (Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994). Support theory proposes that the judged likelihood of any hypothesis is given by the support for that focal hypothesis, as compared with the support for the alternative(s). Judged probability can be thought of as the strength of evidence or support for one hypothesis, relative to the support for its alternatives. Therefore, the judged probability that Hypothesis A will be chosen over B can be given by the equation:

$$P(\text{A} > \text{B}) = \frac{s(\text{A})}{s(\text{A}) + s(\text{B})}. \qquad (1)$$

In practice, the elementary support values are usually assessed by asking judges to *rate* the strength of evidence for each hypothesis, and then those numbers are combined, using Equation 1. For example, Koehler (1996) applied support theory to predictions of the outcomes of 20 NBA basketball games. Participants made independent assessments of the strength of each of the teams and then judged the results of various games. Koehler found that by using these strength assessments as support values and, hence, as inputs to the model, the participants' outcome proba-

bility judgments could be predicted with Equation 1 (he reports r^2 values in excess of .90 for the fits between actual and predicted probability judgments). Similar results have been reported by Fox (1999), who also examined the relationship between assessed strength of evidence and judged probability.

This research shows that a person's predictions are related to conceptions of elementary evidence strengths. But where do the original conceptions of strength come from? What is the relationship between the elementary strength judgment and the actual strength of evidence that exists in the world? One answer to this question is to posit a model for learning strengths from experienced frequencies. Bower and Heit (1992) and Heit, Price, and Bower (1994), following the lead of Estes (1976a), have provided a linear operator model for the learning of strengths. The present research takes a simpler approach to determine the inputs to the judgment process—namely, by applying the relative strength equation to the raw *stimulus* values as inputs. We apply the basic relative strength Equation 1 using either frequencies or proportions to predict choice selections, and argue that the result of a competitive goodness-of-fit test provides evidence for the nature of the underlying input format.

Most theoreticians who have addressed this question have concluded that raw frequencies are likely to be the stimulus that is used to infer strength (Tversky & Kahneman, 1973). However, there has been debate about whether these raw frequencies per se are used or whether they are first converted into more useful relative frequency formats (Brase, Cosmides, & Tooby, 1998; Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995). In this paper, we will refer to absolute frequencies as *frequencies* and to relative frequencies as *probabilities*. However, the reader should keep in mind that our experimental participants estimated proportions of event outcomes or predicted the most likely outcomes of uncertain events. Although these responses may not exactly reflect subjective probabilities or other measures of strength of belief, we believe that all of these measures are interrelated and are probably linear transformations of one another.

Estes (1976a) conducted empirical studies that have provided the most relevant results concerning experimental participants' reliance on frequencies versus probabilities. In his research, he designed an experimental paradigm in which participants were asked to make predictions on the basis of a fictional preference survey. The participants were told that a computer had been programmed to conduct a survey asking polled respondents their preferences for various consumer products (or political candidates). The participants were instructed that they would see the results of a large number of these surveys, after which they would be asked to judge which product they thought was most popular. The set of products about which the participants were queried comprised three independent pairs: A and B, C and D, and E and F. Initially, the participants saw a series of observation trials indicating which of the two products was preferred (by one hypothetical survey respondent) on that particular trial. Fol-

lowing these observation trials, the participants were presented with all product pairs and were asked to choose the more popular member of each pair.

By averaging the results of this binary choice task across products, a judged strength measure was computed for each product, and these strength judgments were then compared with those predicted by a probability or a frequency model. The results indicated that when asked to choose between pairs they had not seen in the observation trials (e.g., A vs. C), the participants chose in line with the absolute frequency of occurrence. In other words, the product that won *most frequently* (e.g., Product C, with 112 wins and a win proportion of .58) was chosen as the most popular, despite the higher relative frequency, or proportion, of wins for the comparison product (Product A, with 90 wins and a win proportion of .62).

In other experiments, Estes fitted mathematical models[1] to account for his findings (Estes, 1972, 1976b, 1987; see also Bower & Heit, 1992) and follow-up experiments showed that the impact of individual trials ("A was preferred to B") on the pairwise judgments diminished with the occurrence of intervening events (Whitlow & Estes, 1979) and that surprising or informative events had a larger impact than did expected events on predictions of occurrence (Neely, 1982).

In the present research, our Experiment 1 was a systematic replication of the Estes (1976a) experiment. In the original experiment, the "winning" product in each pair did not typically win by very much; A beat B 62% of the time, C beat D 58% of the time, and E beat F 54% of the time. We were interested in whether the original result—reliance on frequencies—would be replicated when larger statistical differences discriminated between the two candidates in each pair. This would address one criticism of the original experiment, that the products (candidates) were easily confused with regard to their likelihood to win a trial. Another feature of this replication was a more extensive questioning of the participants in an effort to better understand their judgment strategies.

Estes (1976a) demonstrated that absolute frequency of winning dominates probability or relative frequency of winning, but it is possible that his partially connected stimulus design may have encouraged the participants to guess. Asking a person to judge whether Product A or Product C is more popular, when neither product has been compared directly nor linked indirectly (e.g., via comparisons of performance against a common third opponent), does not support logically valid inferences. Pitz, Englert, Haxby, and Leung (1981) conducted a replication of the original Estes (1976a) experiments, with fewer candidates and with all candidates compared with all others in a round-robin design. Under these conditions, the participants appeared to rely on proportions or relative frequencies to a greater extent than in the original Estes (1976a) experiments. However, Pitz et al. did not systematically manipulate structure; they utilized small values of wins and losses, and they used a declarative format of presentation (information was given to the participant all at once in a summary, symbolic format). Experiment 3 in the pres-

ent series used a method similar to that in Pitz et al., but with experienced events, with larger win–loss frequencies, and with common statistical values being used across stimulus designs.

A comparison of the Estes (1976a) and the Pitz et al. (1981) experimental tasks suggested that the degree of connectedness among the competing candidates has an effect on inferential difficulty and that inferential complexity determines whether participants rely on frequencies or probabilities to make their predictions. More inferences are required to assess relative strength when candidates are not directly connected in the overall stimulus structure; hence, inferential complexity was higher in the original Estes (1976a) experiment than in the Pitz et al. experiment. Our hypothesis was that when given a complete ordering structure, participants would rely on the relative (not absolute) frequencies, since it should be easier for participants to reason about the relative strengths of candidates who were not paired during training.

In the present series, Experiment 1 provided an intermediate level of connectedness: a systematic replication of Estes's (1976a) original stimulus plan with six candidates presented in three isolated pairs. In Experiment 2, we investigated performance with a completely disconnected stimulus set: six candidates presented singly. And Experiment 3 provided participants with a complete transitive structure relating each of the six candidates to their neighbors.

In summary, then, in the present experiments, we examined the effectiveness of a simple relative strength model in predicting human judgments, where the inputs to the model were based on measures of the actual stimuli observed by the participants. We also studied the conditions under which people base their judgments under uncertainty on absolute versus relative frequencies, or proportions. The guiding hypothesis was that when the stimulus structure was fully connected, supporting valid transitive inferences, that participants would rely on relative frequencies. When the structure was incomplete, the participants would rely on absolute frequencies. We used the relative strength model as a combination rule to predict participants' choices and judgments from either absolute frequencies or relative frequencies, and we concluded that the best-fitting model would be diagnostic of the participants' strategy and the input format.

EXPERIMENT 1

Method

Participants. The participants in this experiment were 32 students enrolled in general psychology at the University of Colorado, Boulder. They received partial course credit in exchange for their participation.

Materials. The experiments were conducted using Macintosh computers on which Hypercard software had been installed. A paper-and-pencil questionnaire was also utilized.

Procedure. Before beginning the experiment, the participants were told that their task was to make judgments of several candidates who were competing in an upcoming election. They were informed that six candidates were running in this election, and they

were shown the faces (computer generated) and names of the six: Black, Brown, Gray, Green, Silver, and White. It was pointed out that although the results they saw were real, the particular names and faces were been chosen at random and, therefore, they should not attempt to infer anything about a candidate from his name or face (only male candidates were used in order to keep the materials homogeneous).

Following this, the participants were told that an opinion poll had been conducted prior to the election. This opinion poll was conducted by presenting people with a particular candidate pair and asking the respondent which of these two candidates they would vote for in the election. The participants were told that they would see all of the results of this opinion poll and, on the basis of these results, would be asked various questions. On each observation trial they saw one of three possible pairs of candidates—Black and Brown, Green and Gray, or Silver and White—and then saw the name of the candidate preferred by the respondent in the opinion poll. In order to ensure that the participants did not attend selectively to just the winner of each trial, the participants were required to press a specific key after the presentation of the candidates, depending on which particular pair of candidates appeared. After the correct key was pressed, the winner of that trial was displayed.

All the participants saw 144 observation trials in each of three blocks of trials, with varying frequencies for each pair of candidates. The Black/Brown pairing appeared 33% of the time, the Green/Gray pairing appeared 45% of the time, and the Silver/White pairing appeared 22% of the time. Within the respective pairings, Black won approximately 80% of the time and Brown 20% of the time, Green won 70% of the time and Gray 30% of the time, and Silver won approximately 60% of the time and White 40%. The presentation order of these observation trials was randomized. The participants were not aware of the exact frequencies or probabilities of appearance for the candidates and were explicitly instructed not to count or to take notes on the opinion poll results. See Table 1 for a summary of the design.

Following these observation trials, the participants were informed that they would be asked to predict the results of two-candidate elections. The computer displayed various pairs of candidates, and they were asked to predict which of these two candidates they would expect to win an election (we term this the *pairwise-choice* measure). Each of the 15 possible stimulus pairs was displayed twice, once in each left–right order, again in random sequence. This sequence of 144 observation trials followed by 30 testing trials made up one block. Each participant received three successive blocks of trials. The frequency and win probabilities remained the same in all the blocks.

In addition, after the final block of trials and tests, the participants were given a paper-and-pencil questionnaire and were asked to provide a number of estimates of the stimuli presented in the course of the experiment. They were asked to make judgments regarding outcomes of paired elections again. They judged all possible pairs of candidates and estimated the percentage of the vote that each candidate in the pair would get in an election between the two (the *pairwise-proportion* measure). They further estimated the percentage of the vote each candidate would get in a full six-way election (the *full-election* measure). Finally, they were asked to give a brief written report of how they had made their judgments.

Table 1
Design of Experiment 1

Pair	Candidates (Win Probabilities; Win Frequencies)	Presentation Frequency
1	Black (.80; 38) versus Brown (.20; 10)	48
2	Green (.70; 45) versus Gray (.30; 19)	64
3	Silver (.60; 19) versus White (.40; 13)	32

Results

Analyses were conducted on the participants' responses to both the computer and the questionnaire sections of the experiment. A reason for administering the trials in three separate blocks was to look for any possible effects of learning over the course of the experiment; however, examination of the results showed no significant changes between blocks, and therefore, for the analyses reported here, results have been collapsed across blocks. Three measures were the focus of our analyses: the pairwise choices made via computer, the pairwise vote proportions from the questionnaire, and the six-way full-election vote proportions, also from the questionnaire. Initial analyses of these measures indicated that all three were highly intercorrelated [full-election–pairwise-proportion, $r(5) = .99$, $p < .001$; full-election–pairwise-choice, $r(5) = .97$, $p < .001$; pairwise-choice–pairwise-proportion, $r(5) = .99$, $p < .001$].

Pairwise comparisons, group average results. The primary use of the strength model was to predict the results of individual pairwise elections. The dependent variables in this analysis were the judgments of paired elections made as part of the postexperimental questionnaire (pairwise-proportion measure). All 15 pairs were shown, and the participants allocated a percentage of the vote to each candidate in the pair, so that each pair totaled 100%. These data are in the form of proportions, and so a log-odds transformation of the data was computed. However, since the vast majority of the values lay between .2 and .8, this transformation had little effect, and hence, the results reported were based on the untransformed data. Two different predictions of pairwise elections were made by participants—one in the computer portion of the experiment, in which the pairs were displayed and the participants merely had to pick the likely winner, and one in the questionnaire portion, in which pairs were shown and the participants had to indicate the proportion of the vote each candidate would likely get. These two judgments were highly correlated across candidate pairs [$r(15) = .98, p < .001$]; therefore, in the interest of brevity, the following statistical analyses will primarily concentrate on the responses given in the questionnaire. Analyses were also conducted on the computer responses, and no significant deviations were found from those of the questionnaire responses.

The strength model was used to predict these overall candidate vote percentages on the basis of the stimuli seen by the participants. The candidates' frequency of winning in the training trials was the input to the model (this was the *strength measure* used for each candidate). The pairwise model took the form

$$P(A > B) = \frac{s(A)}{s(A) + s(B)}$$

(Equation 1, above)

where $s(A)$ is the overall win frequency of Candidate A and $s(B)$ the overall win frequency of Candidate B (note

that we assume that the model's predicted probability and the participants' proportion of win choices are directly related).

The simple strength model based on win frequency did quite well at predicting the mean judgments of the participants, yielding an R^2 of .88 (RMSD = 7.20, SE = 2.31).[2] By means of comparison, models based on candidates' overall appearance frequency (i.e., the absolute number of appearances of each candidate in the training trials) and candidates' overall win probability (i.e., the proportion of times the candidate won during training) were also constructed. Both the appearance model ($R^2 = .22$) and the probability model ($R^2 = .55$, RMSD = 11.53) exhibited fits inferior to those for the win frequency model. Figure 1 plots the predictions for the frequency and probability models versus the vote percentage awarded to each of the candidates, averaged across all 15 pairs for ease of display.

Six-candidate election results. The simple strength model was used to predict the estimates of percentages of votes for each candidate in a six-way election. The strength model was used to predict these overall candidate percentages on the basis of the stimuli seen by the participants. For example, when attempting to predict the overall success of Green, we used the model $P(\text{Green}) = s(\text{Green})/[s(\text{Green}) + s(\text{Black}) + s(\text{Brown}) + s(\text{Gray}) + s(\text{Silver}) + s(\text{White})]$, where $s(X)$ is the win frequency or win proportion of each candidate (X); and we assumed that the predicted probability and the estimated percentage of the vote were directly related.

The model based on the candidates' frequency of winning was a significant predictor of the observed data [$F(1,5) = 56.99, p < .01$], with an R^2 value of .93, indicating that the model could account for almost all of the variance in the participants' judgments. A model based on candidates' probability of winning also proved a significant predictor [$F(1,5) = 18.79, p < .05$], although it accounted for a smaller proportion of the variance ($R^2 = .82$).

Individual participant strategies. Although the models tested were intended to simulate the mean performance of a group of participants and contained no parameters allowing differential predictions for different people, it is of interest to examine how the model can account for individual responses. Two approaches were used to identify the combination rules used by individual participants: a traditional cluster analysis on the raw percentage estimates (input to the cluster analysis were the pairwise percentage estimates given in the questionnaire) and a coefficient W as an index of the degree to which an individual's pattern of judgments depended on frequencies versus proportions (assuming that the relative strength model was the integration rule; see the Appendix for the derivation of this coefficient).

The analyses converged to a remarkable degree, and the participants were sorted into two groups: frequentists ($n = 22$, mean $W = -.14$) and probabilists ($n = 8$, mean $W = .73$). The sorting was made on the basis of the output from the cluster analysis, and the groups differed significantly in terms of their W values [$t(30) = -5.02, p < .01$]. There
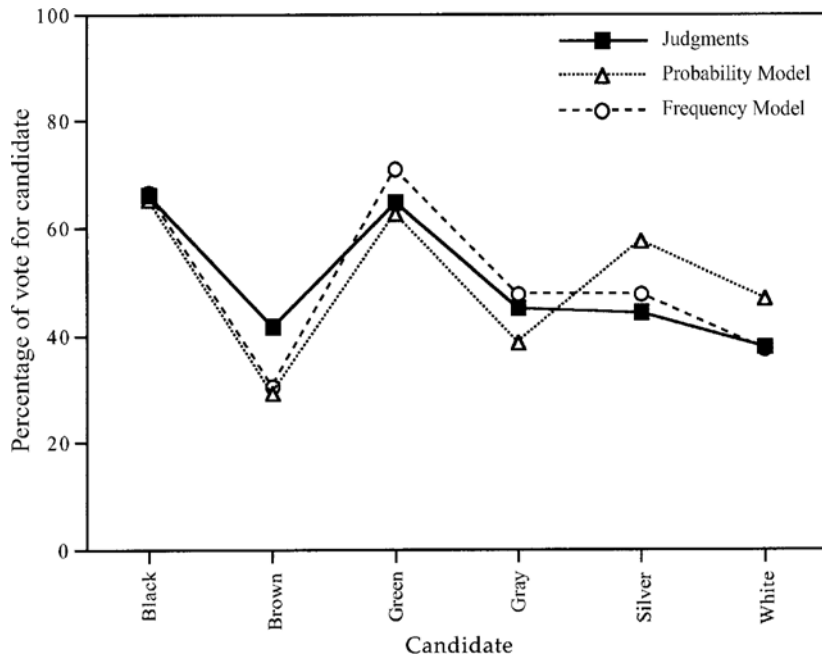
Figure 1. Experiment 1: Judgments versus models.

was a good fit between the frequency model and the frequentists' judgments ($R^2 = .80$); the frequentists' fit to the probability model was much lower ($R^2 = .39$). In addition, the probability-model–probabilists fit was good ($R^2 = .88$), although the probabilists also had quite a high fit to the frequency model ($R^2 = .72$).

Discussion

In summary, Experiment 1 provided a good initial test of the usefulness of a simple strength model based on actual stimuli as a predictor of pairwise judgments. This result confirms Estes's (1976a) original conclusion that win frequency, and not probability, is used by participants as the basis for judgments when the stimulus events are disconnected. That is, in Estes's stimulus design, only a subset of the pairwise relationships were displayed in the empirical event structure, a situation that we hypothesize would induce thinking about the candidates in terms of their win frequencies. To further test this hypothesis—namely, that a disconnected set of candidates would lead to use of win frequency—in Experiment 2, we examined the effect of a completely independent stimulus set and asked the question of what would happen when the candidates were displayed individually during training, as opposed to in pairs.

Further analyses at the individual level revealed that the participants exhibited two distinct judgment strategies— we labeled them *frequentists* and *probabilists*—reflecting the aspect of the stimulus that was relied on to estimate candidate strength or propensity to win an election. Although most of the participants appeared to follow the relative strength integration rule summarized in the strength

equation, approximately 75% of the participants relied on frequency, the remaining 25% using proportion of wins as inputs to the model.

Use of the model to predict overall candidate judgments also indicated that a model based on win frequency could accurately predict overall candidate judgments. In this analysis, however, a model based on win probability also proved a significant predictor of judgments, a result we attribute to the high correlation between the predictions based on frequencies and the predictions based on probability (as was also the case with the original Estes, 1976a, stimuli, where the correlation between win frequency and win probability was $r = .56$). A further goal, therefore, for Experiment 2 was to create a stimulus set that would increase independence between the win frequencies and the probabilities of the candidates.

EXPERIMENT 2

Method

Participants. The participants in this experiment were 39 students enrolled in general psychology at the University of Colorado, Boulder. They received partial course credit in exchange for their participation.

Materials. The materials were the same as those in Experiment 1.

Procedure. The experimental paradigm was the same as that in Experiment 1—namely, a simulation of an upcoming election. The participants were told that six candidates were running in the election and, as before, were shown pictures and names of the candidates: Brown, Green, Orange, Red, Silver, and White. Different candidate name–face combinations were used to eliminate any potential effect of biases owing to these characteristics. Again, they were told that an opinion poll had been conducted on pairs of candidates. In

Table 2
Experiments 2 and 3: Candidates' Overall Win
Probability and Frequency

Candidate	Trials	Wins (Win Frequency)	Losses	Win Probability
Red	20	15	5	.75
Brown	70	43	27	.61
Orange	80	35	45	.44
Green	90	52	38	.58
Silver	100	45	55	.45
White	40	10	30	.25

this study, they were informed that they would see only one of the two candidates on each trial and, then, see whether that candidate had been preferred.

On each observation trial they saw one of six possible candidates—Red, Brown, Orange, Green, Silver, or White—and then saw whether the respondent in the opinion poll would vote for the candidate or not. Each participant saw 400 observation trials, with varying frequencies— both of appearance and winning—for each candidate. The presentation order of these observation trials was randomized. The participants were not aware of the exact frequencies or probabilities of appearance for the candidates and were again explicitly instructed not to count or to take notes on the opinion poll results. See Table 2 for a summary of the design.

Following these training trials, the participants were tested as described in Experiment 1. Both computer and paper-and-pencil tests were administered, asking the participants to judge the winner of each candidate–candidate pair election, as well as judgments of a full six-way election and a brief written description of their judgment process.

Results

Analyses were again conducted on the participants' responses to both the computer and the questionnaire sections of the experiment. The participants made three separate

judgments: predict the winner from each pair, predict the percentage of votes for both candidates in each pair, predict the percentage of votes for each candidate in a full six-way election. And as in Experiment 1, the three basic judgment measures were highly intercorrelated [full-election–pairwise-proportion, $r(5) = .96, p < .001$; full-election-pairwise-choice, $r(5) = .94, p < .001$; pairwise-choice–pairwise-proportion, $r(5) = .95, p < .001$].

Pairwise comparisons, group average results. In this experiment, as in the previous one, the primary use of the strength model was to predict the results of the individual pairwise elections. The dependent variables were the same as those outlined in Experiment 1 (the pairwise-proportion measure). As before, a log-odds transformation was computed but had little effect on the results, and so the results reported were again based on the untransformed data. The two separate predictions of pairwise elections made by participants—one in the computer portion of the experiment and one in the questionnaire portion—were highly correlated across candidate pairs [$r(15) = .98, p < .001$], and so again we will concentrate statistically on the estimates of percentages of votes given in the questionnaire.

The relative strength model was used to predict these overall candidate percentages on the basis of the stimuli seen by the participants. The first model tested was of the same type as the one that had had the best fit in the previous experiment—namely, one based on the candidates' win frequency. The pairwise model again took the form of Equation 1. Figure 2 plots the predictions for the frequency and probability models versus the percentage of votes awarded to each of the candidates, averaged across all 15 candidate pairs.
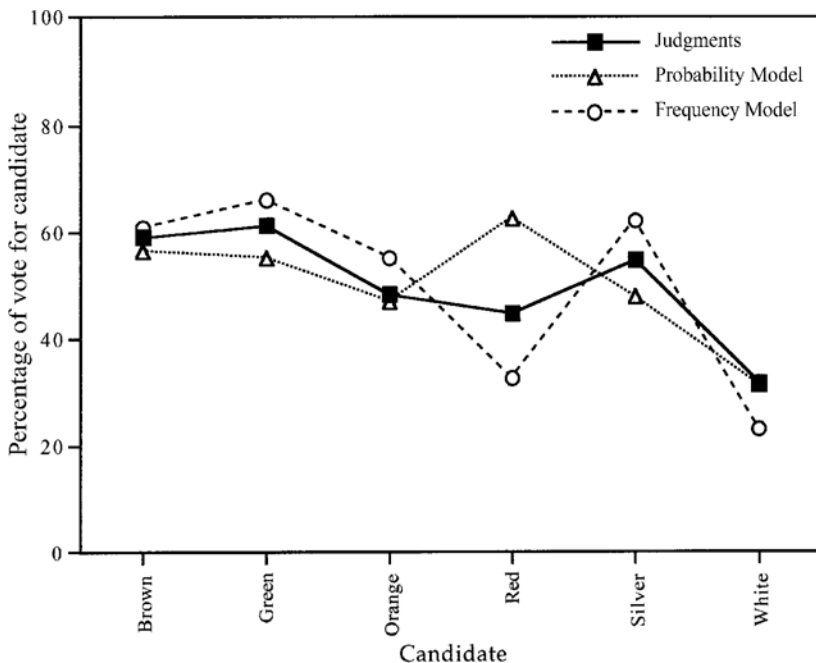

Figure 2. Experiment 2: Judgments versus models.

Table 3
Experiment 2: Candidate Pairs With Differential Predictions by Frequency and Probability

High-Probability Candidate	Selection %	High-Frequency Candidate	Selection %
Red*	37	Green	63
Brown	48	Green	52
Brown	54	Silver	46
Red*	37	Brown	63
Red	46	Orange	54
Red*	40	Silver	60

*$p < .05$.

The strength model based on win frequency again had quite a high fit when predicting the mean judgments of the participants (R^2 of .89, RMSD = 9.05, SE = 2.81), whereas once more the simple strength model whose input was the overall win probability of the candidate did more poorly at predicting judgments (R^2 = .39, RMSD = 10.92). We believe this result to again demonstrate that with a disconnected set of stimuli, participants will have a tendency to rely on the absolute frequency of wins achieved by each candidate.

A final qualitative test of whether frequency or probability is the basis for judgments can be found by asking what happens when two candidates are paired together, one of whom has a higher probability of winning than the other, when the other has a higher frequency of winning. Six of these pairs existed, and their results are summarized in Table 3. In five of these six cases, the high win frequency candidate was selected more often than the high win probability candidate (although this was statistically significant only in three of the cases).

Six-candidate election results. As in Experiment 1, the generality of the simple strength model was tested by attempting to predict the estimates of percentages of votes in a six-way election. The primary dependent variables were computed exactly as in Experiment 1. Since all of these measures again proved highly intercorrelated (shown above) and once more yielded essentially the same results, in this section we will utilize the responses given when the participants were asked to apportion the vote out among all six candidates. The type of strength model used was the same as that used to calculate global predictions in Experiment 1—that is, $P(\text{A wins}) = s(\text{A})/[s(\text{A}) + s(\text{B}) + s(\text{C}) + s(\text{D}) + s(\text{E})]$.

The model based on the candidates' frequency of winning was a significant predictor of the observed data [$F(1,5) = 16.49, p < .02$], with an R^2 value of .80, indicating that the model could account for a substantial portion of the variance in the participants' judgments. A model based on the candidates' win probability was not a significant predictor [$F(1,5) = 1.91, p > .05; R^2 = .32$]. This performance demonstrates that a model using the candidates' win frequency is again a better predictor of participants' judgments.

Individual participant strategies. We again used two approaches to attempt to identify the inputs used by indi-

vidual participants: a cluster analysis in which the pairwise percentage estimates was used and the W coefficient. Two groups were identified by the cluster analysis, and the members of each group generally had W values very similar to each other. We again labeled the groups frequentists ($N = 22$, mean $W = 0.14$) and probabilists ($N = 16$, mean $W = 0.93$). The W values across groups differed significantly [$t(36) = -7.29, p < .01$]. The frequency model had a better fit to the frequentists' average judgments ($R^2 = .92$) than did the probability model to this group's judgments ($R^2 = .03$). When the probabilists were assessed, the probability model showed a closer fit to the probabilists' judgments ($R^2 = .95$) than to the frequentists' ($R^2 = .26$).

Discussion

Experiment 2 provided a further test of the strength model and showed once again that a simple model based on stimulus information can make accurate predictions of judgments under uncertainty. As in Experiment 1, the inputs to the best-fitting model were absolute win frequencies, a result we attribute to the disconnected structure of the stimulus set. Individual subject analyses gave further credence to this hypothesis. Use of different color names for the stimuli across Experiments 1 and 2 also demonstrated that there appeared to be no effect of the specific color names or faces used as stimuli. Also of interest was the sorting of participants into two groups: those who appeared to use the candidates' win frequency and those who, instead, appeared to be influenced by their win probability.

Both of the experiments described have used, to some degree or other, a disconnected stimulus set—that is, only a subset of the pairwise relationships were expressed in the empirical event structure. Experiment 3 extended these results by exploring the effect of providing a more complete, transitive, interevent structure in the to-be-judged stimulus materials, while keeping overall frequency and probability rates identical.

EXPERIMENT 3

Method

Participants. The participants in this experiment were 36 students enrolled in general psychology at the University of Colorado, Boulder. They received partial course credit in exchange for their participation.

Materials. The materials were the same as those in Experiment 1.

Procedure. The experimental paradigm was the same as those in both previous experiments. The participants were told that six can-

Table 4
Design of Experiment 3

Pair	Candidates (Win Probabilities; Win Frequencies)	Presentation Frequency
1	Red (.75; 15) versus Brown (.25; 5)	20
2	Brown (.75; 38) versus Orange (.25; 12)	50
3	Orange (.75; 23) versus Green (.25; 7)	30
4	Green (.75; 45) versus Silver (.25; 15)	60
5	Silver (.75; 30) versus White (.25; 10)	40

didates were running in the election—Brown, Green, Orange, Red, Silver, and White—and were shown pictures and names of the candidates. They were told that an opinion poll had been conducted and that they would see the results of that polling.

On each observation trial, the participants saw one pair of these candidates and then saw which candidate a respondent in the opinion poll had preferred. Each participant saw 200 trials, with the presentation order of the testing trials being randomized. The details of the pairings shown to the participants are given in Table 4. The overall win probabilities and frequencies (collapsed across all paired trials) for each of the six candidates were identical to those used in Experiment 2 (see Table 2 for these overall probabilities and frequencies). The only difference was, of course, that the candidates in this experiment were displayed in pairs, not individually (and thus there were half as many trials).

Following these training trials, the participants were tested as described in the previous experiments, with both computer and paper-and-pencil tests being administered. Furthermore, they also gave estimates of how many times each pair appeared during training and what proportion of these trials was won by each candidate of the pair.

Results

Analyses were conducted on the participants' responses to both the computer and the questionnaire sections of the experiment. Once again, as in the previous experiments, the three basic judgment measures were highly intercorrelated [full-election–pairwise-proportion, $r(5) = .94, p < .001$; full-election–pairwise-choice, $r(5) = .93, p < .001$; pairwise-choice–pairwise-proportion, $r(5) = .99, p < .001$].

Judgment accuracy. As a preliminary analysis, the quality of the participants' judgments was examined by assessing the accuracy of the frequency estimates they were required to make upon completion of the experiment. Statistical analysis demonstrated that, except in one case, there were no significant differences between the estimated and the actual number of times each pair appeared in training [Br/Or (actual = 50, mean estimate = 48.31), $t(35) = -0.57, p > .05$; Or/Gr (actual = 30, mean estimate = 30.03), $t(35) = 0.03, p > .05$; Rd/Br (actual = 20, mean estimate = 20.64), $t(35) = 0.54, p > .05$; Gr/Si (actual = 60, mean estimate = 56.25), $t(35) = -3.75, p < .05$; Si/Wh (actual = 40, mean estimate = 41.17), $t(35) = 0.50, p > .05$]. The correlation between actual and estimated appearances for the pairs was $r(5) = .99, p < .01$.

The participants were also asked to estimate how many times each candidate appeared in total during the training trials. Once again, the estimates proved very accurate, yielding a correlation between the two measures of $r(5) = .92, p < .01$.

Pairwise comparisons, group average results. In this experiment, as in the previous two, the primary use of the strength model was to predict the results of the individual pairwise elections. The dependent variables were the same as those outlined in the previous experiments (a log-odds transformation was again computed but had no effect). The two separate predictions of pairwise elections made by participants—one in the computer portion of the experiment and one in the questionnaire portion—were again highly correlated across candidate pairs [$r(15) = .99, p < .001$]. We will discuss the estimates of percentages of votes given in the questionnaire.

As was hypothesized, this time the strength model based on win frequency did not do as well at predicting the mean judgments of the participants (R^2 of .31, RMSD = 17.49, $SE = 3.41$)—primarily, we believe, owing to the inferentially fully connected stimulus set in this experiment. As confirmation, the simple strength model, whose input was the overall win probability of the candidate, had a much higher fit predicting the participants' judgments ($R^2 = .81$, RMSD = 6.45). Figure 3 plots the models' predictions versus the participants' judgments for each of the candidates, demonstrating the closer fit of the probability model.

The simple strength model being tested here implies that for each pairwise judgment, the participants were, in some sense, computing a single *propensity* for each candidate and using it in their estimates. Presumably, this propensity was calculated by using information from all of the trials that the particular candidate appeared in. But what happens when a participant is presented with a pair that has already been seen in training? Does he or she use this inferred propensity? Or does he or she, instead, simply attempt to remember what actually happened in the training trials? A further test of the model therefore concerns the judgments of the five pairs that the participants had already seen in training. Figure 4 plots the actual proportion of the time that the first-named candidate of each pair won, the proportion allocated by the participants themselves, and the proportions predicted by models based on both total win probability and total win frequency (across all pairs). As can be seen, the win probability model was a significant predictor of the participants' responses [$F(1,4) = 35.38, p < .01$], in contrast to predictions based on the actual stimuli the participants had previously seen. The model based on win frequency was not a significant predictor of the participants' judgments [$F(1,4) = 2.93, p > .05$].

A final qualitative test of whether frequency or probability is the basis for judgments can be found by asking what happens when two candidates are paired together, one of whom has a higher probability of winning than the other, but the other of whom has a higher frequency of winning. Six of these pairs existed, and their results are summarized in Table 5. In direct contrast to the results obtained in the previous experiment—which used exactly the same stimulus numbers—out of the six cases, the high win probability candidate was selected more often than the high win frequency candidate on five occasions (although it should be noted that only three of these six cases differed significantly).

Six-candidate election results. Once again, we attempted to predict the estimates of percentages of votes in a six-way election, in an effort to test the generality of the simple strength model. The primary dependent variable was again the individual candidate ratings averaged across all of the pairwise comparisons.

The type of strength model used was the same as that used to calculate global predictions in Experiment 1. The model based on the candidates' probability of winning was a significant predictor of the observed data [$F(1,5) = 16.70, p < .02$], with an R^2 value of .96, indicating that the
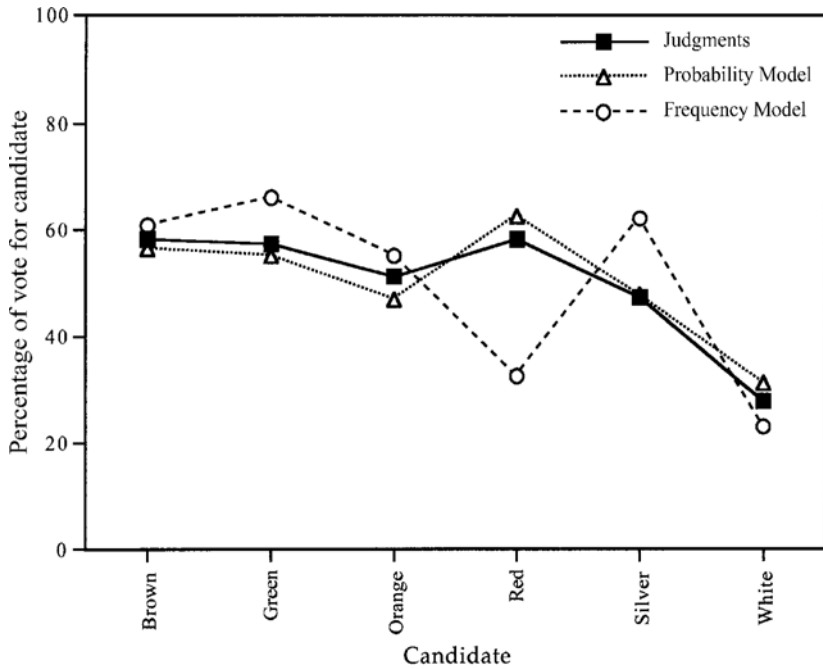
Figure 3. Experiment 3: Judgments versus models.

model could account for a substantial portion of the variance in the participants' judgments. A model based on candidates' win frequency was not a significant predictor [$F(1,5) = 1.78, p > .05; R^2 = .06$].

Individual participant strategies. As in Experiments 1 and 2, two approaches were used to identify the combination rules used by individual participants: a cluster analysis on the pairwise percentage estimates and Norman An-
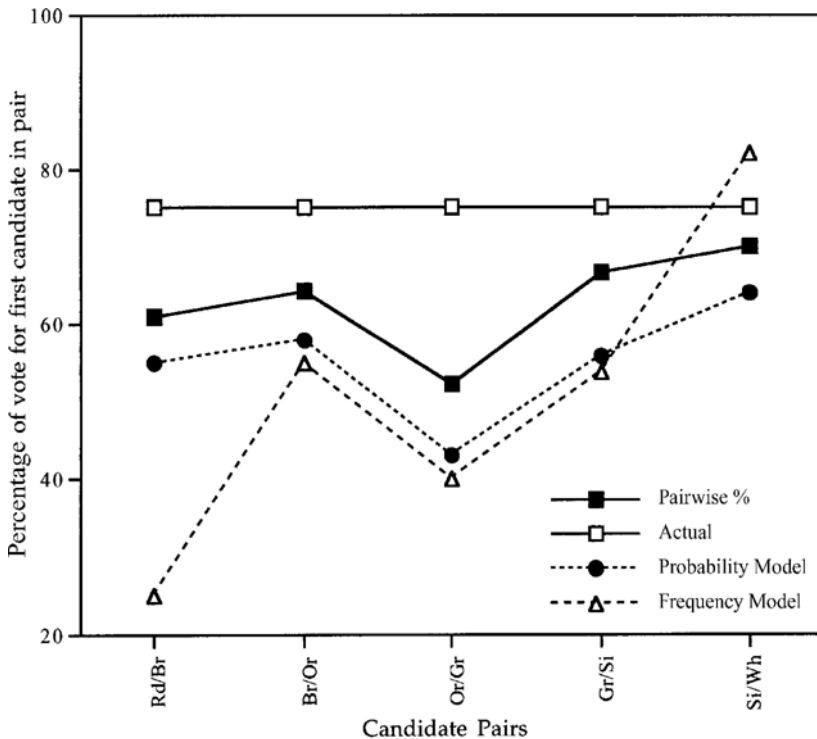

Figure 4. Experiment 3: Judgments and models' predictions for training data. Rd, Red; Br, Brown; Or, Orange; Gr, Green; Si, Silver; Wh, White.

Table 5
Experiment 3: Candidate Pairs With Differential Predictions by Frequency and Probability

High-Probability Candidate	Selection %	High-Frequency Candidate	Selection %
Red	48	Green	52
Brown	53	Green	47
Brown*	63	Silver	37
Red*	67	Brown	33
Red*	61	Orange	39
Red	54	Silver	46

*$p < .05$.

derson's coefficient W. Again, the analyses generally converged, with three groups emerging from the cluster analysis: frequentists ($n = 13$, mean $W = 0.12$), probabilists ($n = 22$, mean $W = 1.20$), and unclassified ($n = 1$). The W values across the two groups of interest were significantly different [$t(33) = -8.30$, $p < .01$]. The frequency model matched the frequentists' average judgments ($R^2 = .88$) much better than the probability model ($R^2 = .08$); the probability model had a substantially closer fit to the probabilists' judgments ($R^2 = .85$) than to the frequentists' ($R^2 = .04$).

A cursory examination of the written protocols completed by all the participants, although not systematically analyzed, found that those participants labeled as *probabilists* from the statistical analyses were more likely to report having consciously relied on relative probabilities in making their judgments, whereas the *frequentists* based their judgments on how many wins each candidate had.

Discussion

Experiment 3 provided a further test of the strength model in predicting quantitatively the participants' judgments. The most interesting finding was that the inputs to the best-fitting model were relative frequencies or probabilities, rather than absolute frequencies, as in the two previous experiments. Of particular interest was the fact that probabilities were here a better fit even though the stimulus win probabilities and frequencies presented in this experiment were identical to those of the previous experiment. We attribute this change to the provision of a stimulus set structure in which transitive orders of *triples* of candidates were readily perceptible.

GENERAL DISCUSSION

The present research explores the cognitive sources of primitive confidence or strengths of belief about the occurrences of simple, uncertain events. The experiments were designed to create a situation in which participants were forced to rely on the statistical properties of the environment to make judgments under uncertainty and, as such, effectively ruled out any inferences except those based on frequencies, proportions, or other statistical properties of the poll reports.

Within this simple task environment, we attempted to identify the statistical properties of the data that the par-

ticipants relied on when they made their judgments. Specifically, the election judgment task was designed to allow the separation of frequency and probability as potential mediators of judgments of likelihood. The first experiment demonstrated that when the candidate pairings in the polls are piecemeal, so that there is no fully connected relational structure in the data, most people rely on absolute win frequencies to predict who will win and by how much. The second experiment eliminated all interevent structure from the to-be-judged events and showed that most of the participants used win frequencies. The win frequency strength model again accurately predicted the participants' responses. The third experiment showed that when the stimulus set was structured in a connected manner, so that the participants could more easily make transitive inferences between any two candidates, most participants relied on win probabilities, rather than on win frequencies. Of note is that whereas in these latter two experiments the candidates possessed identical win frequencies and probabilities, nonetheless there was a clear switch from frequency to probability as the interevent statistical structure underlying the poll results became more coherent. The more interevent relationships that were experienced in the data, the more likely were the participants to rely on probabilities, rather than on frequencies.

The critical aspect of statistical structure, in the present experiments, is captured by the notion of *relational complexity*—that is, the number of variables that can be related to one another within a single representation (Halford, Wilson, & Phillips, 1998). In the present judgment task, our interpretation is that when people are able to make inferences about relative frequencies, they will do so, inferring these higher order statistical properties from more primitive experienced frequencies. We hypothesize that owing to the greater number of variables that must be processed in parallel in the relationally demanding structures (i.e., Experiments 1 and 2), the more computationally exacting aspects of the stimuli are increasingly difficult to calculate and maintain, and hence participants are forced to rely on simpler representations (namely, absolute frequencies). When the inferential task appears to be difficult or even impossible (because of limited interevent relational information in some of the present experiments), participants will rely on the primitive raw frequency information. However, when multiple variables can be processed together, as occurred in the connected set of Experiment 3, there are more cognitive resources left to attempt to capture more complex features of the to-be-judged stimuli, such as relative frequencies. The general principle underlying this interpretation is that when a task requires a greater number of inferences, people revert to simpler, more primitive strategies to respond. The emphasis on information representation as a fundamental part of a theory of judgment is becoming more popular as research on judgment becomes more "cognitive" (Hastie, 2000). Recent research on reasoning about probability problems confirms that the nature of the solver's mental representation of the structure of a set of to-be-judged events, created during the comprehension of the original

problem statement, has dramatic effects on the quality of the final solution (e.g., Fiedler, 2000; Girotto & Gonzalez, 2001; Johnson-Laird, 1994; Lovett & Schunn, 1999; Nickerson, 1996).

Like many other researchers, we presume that although absolute frequency of occurrence information is elementary and may even be the automatic product of a "hard-wired" event counter (cf. Hasher & Zacks, 1984; Hintzman, Nozawa, & Irmscher, 1982; although controlled strategies can also be employed deliberately to estimate frequencies—see, e.g., Brown, 1995; Jonides & Naveh-Benjamin, 1987), the inference of relative frequencies, proportions, or probabilities requires additional, possibly controlled strategies (see Sedlmeier, 1999). It is tempting to conjecture that the two behavioral models, frequentist versus probabilist, are associated with two distinct memory/cognitive systems (Sloman, 1996; Smith & DeCoster, 2000). The hypothesis would be that the frequentist model describes the outputs of a slow-learning, associationist, automatic cognitive process, whereas the probabilist model describes a fast-learning, symbolic, controlled process (McClelland, McNaughton, & O'Reilly, 1995).

Throughout the present experiments, we observed substantial, systematic, and readily interpretable individual differences in judgment strategies. The present methods allowed us to classify individual participants as frequentists or probabilists according to individual strategies, and indeed, the majority of the participants could be typed as one or the other. An obvious question to ask is what makes some participants probabilists and some frequentists. Although the present study does not address this question, some ideas may be gleaned from recent research on individual differences in reasoning (for a review, see Stanovich & West, 2000). Those results suggest that some kind of general cognitive ability, or at least the aptitude to perform conventional scholastic tasks, may influence performance and strategy selection in judgment and reasoning tasks. Again, without the relevant data, we can only speculate, but it seems plausible that cognitive ability may be one factor that distinguishes between frequentists and probabilists in the present experiments.

A collateral conclusion of the present research is an endorsement of the relative strength combination rule for integrating strengths of belief, propensities to occur, or degrees of support in judgment under uncertainty. The relative strength principle is central in major theories of judgment under uncertainty—most notably, support theory, information integration theory, and Luce's (1959) choice axiom. Our finding that there was a high degree of consistency between the participants' judgments of outcomes in multicandidate (six-way) elections and on the pairwise election trials provides further support to the claim that the participants had a unitary sense of the "strength" of each candidate, a strength that was largely independent of the means by which it was inferred.

One open question, which is highlighted by the present research, concerns the nature of a normative solution for the pairwise election judgment tasks. To the best of our knowledge, there is no generally accepted normative rule for predicting the probability of a win in a pairwise election from individual survey results, especially when no direct comparisons are available. Nor is there an obvious normative answer for the estimation of percentage of votes in a six-way election. Despite this lack of "correct" solutions, the participants did seem to converge on answers to the judgments on the basis of frequencies or on probabilities.

The most important results of the present research concern the variation in individual strategies, with respect to the statistical aspects of the experienced events in judgments under uncertainty. The participants in these experiments appeared to optionally rely on frequencies or proportions when inferring propensities to win political elections. We identified one of the task conditions that determines which strategy is followed: When the set of to-be-judged events is coherent and well connected—for example, in a complete and consistent transitive ordering—people appreciate this structure and rely on higher order statistical properties, such as relative frequencies and probabilities. But when the events lack a coherent global structure, people rely on more primitive stimulus properties, such as absolute frequencies of occurrence. This conclusion identifies an important interaction between statistical information, the cognitive representation of interevent relationships, and judgment strategies.

REFERENCES

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

BOWER, G. H., & HEIT, E. (1992). Choosing between uncertain options: A reprise to the Estes scanning model. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honor of William K. Estes* (Vol. 1, pp. 21-43). Hillsdale NJ: Erlbaum.

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

BROWN, N. R. (1995). Estimation strategies and the judgment of event frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1539-1553.

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

ESTES, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, **67**, 81-102.

ESTES, W. K. (1976a). The cognitive side of probability learning. *Psychological Review*, **83**, 37-64.

ESTES, W. K. (1976b). Some functions of memory in probability learning and choice behavior. *Psychology of Learning & Motivation*, **10**, 1-45.

ESTES, W. K. (1987). Application of a cognitive-distance model to learning in a simulated travel task. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 380-386.

FIEDLER, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, **107**, 659-676.

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

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

GIROTTO, V., & GONZALEZ, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, **78**, 247-276.

HALFORD, G. S., WILSON, W. H., & PHILLIPS, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental and cognitive psychology. *Behavioral & Brain Sciences*, **21**, 803-864.

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

HASTIE, R. (2000). Problems for judgment and decision making. *Annual Review of Psychology*, **52**, 653-683.

HEIT, E., PRICE, P. C., & BOWER, G. H. (1994). A model for predicting the outcomes of basketball games. *Applied Cognitive Psychology*, **8**, 621-639.

HINTZMAN, D. L., NOZAWA, G., & IRMSCHER, M. (1982). Frequency as a nonpropositional attribute of memory. *Journal of Verbal Learning & Verbal Behavior*, **21**, 127-141.

JOHNSON-LAIRD, P. N. (1994). Mental models and probabilistic thinking. *Cognition*, **50**, 189-209.

JONIDES, J., & NAVEH-BENJAMIN, M. (1987). Estimating frequency of occurrence. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 230-240.

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

LEON, M., & ANDERSON, N. H. (1974). A ratio rule from integration theory applied to inference judgments. *Journal of Experimental Psychology*, **102**, 27-36.

LOVETT, M. C., & SCHUNN, C. D. (1999). Task representations, strategy variability, and base-rate neglect. *Journal of Experimental Psychology: General*, **128**, 107-130.

LUCE, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York: Wiley.

MCCLELLAND, J. L., MCNAUGHTON, B. L., & O'REILLY, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419-457.

NEELY, J. H. (1982). The role of expectancy in probability learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **8**, 599-607.

NICKERSON, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin*, **120**, 410-433.

ODEN, G. C., & LOPES, L. L. (1997). Risky choice with fuzzy criteria. *Psychologische Beiträge*, **39**, 56-82.

PITZ, G. F., ENGLERT, J. A., HAXBY, K., & LEUNG, L. S. (1981). Learning conditional frequencies in a probability learning task. *Acta Psychologica*, **47**, 229-243.

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

SEDLMEIER, P. (1999). *Improving statistical reasoning: Theoretical models and practical implications.* Mahwah, NJ: Erlbaum.

SLOMAN, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, **119**, 3-22.

SMITH, E. R., & DECOSTER, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality & Social Psychology Review*, **4**, 108-131.

STANOVICH, K. E., & WEST, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral & Brain Sciences,* **23**, 645-726.

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.

WHITLOW, J. W., JR., & ESTES, W. K. (1979). Judgments of relative frequency in relation to shifts in event probability: Evidence for a limited capacity model. *Journal of Experimental Psychology: Human Learning & Memory*, **5**, 395-408.

WICKENS, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, **105**, 379-386.

NOTES

1. Estes proposed a model for predicting choices [e.g., $P(A,B)$, or the probability Candidate A will be chosen over B from a choice set containing A and B] as a function of the frequencies of wins (W_A, W_B) and losses (L_A, L_B) experienced by the participant (1976a, p. 58, Equation 9). This model incorporated three free parameters and additional terms including wins, losses, both, and a ratio form. As it turns out, the simple, "zero parameter" ratio model (Equation 1) yields fits to the grouped data that are statistically indistinguishable from Estes's three free parameter model; the simple ratio model fits slightly better in Experiment 1, and Estes's model fits slightly better in Experiments 2 and 3. Since our goal in the present analysis is to pit absolute frequency against relative frequency at the individual participant level, we compare the two versions of the simple ratio model, since this approach provides equally good fits overall and much more transparent conceptual interpretations.

2. Note that the values of R^2 reported here are squared Pearson correlations of observed and predicted values. We also fitted models based on other variants of R^2, such as the fitted sums of squares relative to the sum of squares about the mean [$R^2 = 1 - \Sigma(y_i - x_i)^2 / \Sigma(y_i - y_{bar})^2$; see Wickens, 1998]. The results across measures were highly consistent—for example, in Experiment 1 the R^2 fit between the observed data and the frequency model was .81, whereas the R^2 fit between the observed data and the probability model was .38, demonstrating a similar qualitative pattern of results. For ease of explanation, the standard R^2 measure will be used throughout the rest of this paper.

APPENDIX
Norman Anderson's Derivation of the *W* Statistic

Consider the data for a single participant.

Let Y be the participant's score in a given cell of the triangular design.

Let x be the score predicted from probability(win) strengths.

Let z be the score predicted from frequency(win) strengths.

Suppose that the participant's response is an average of the two strength rules: $Y = Wx + (1-W)z$. (Or equivalently, that one rule is followed with probability W, the other rule with probability $1 - W$.)

We seek to estimate W by least squares, minimizing the sum of squared deviation, summed over the 15 cells of the triangular design for the given subject:

$$SS = \sum_{\text{cells}} [Y - Wx - (1-W)z]^2 = \text{minimum}.$$

Differentiate SS with respect to W, set equal to 0, and solve for W:

$$\frac{dSS}{dW} = 2\sum [Y - Wx - (1-W)z](-x+z) = 0.$$

$$\sum Y(z-x) = \sum [Wx + (1-W)z](z-x)$$

$$= W\sum -(z-x)^2 + \sum [z^2 - zx].$$

$$W = \frac{\sum (x-z)(Y-z)}{\sum (x-z)^2}.$$

This gives W on the left in terms of known numbers on the right. If the participant follows one rule exclusively, then W is 0 or 1.