Subjective Probability: A Judgment of Representativeness

DANIEL KAHNEMAN AND AMOS TVERSKY[1]

The Hebrew University, Jerusalem

This paper explores a heuristic—*representativeness*—according to which the subjective probability of an event, or a sample, is determined by the degree to which it: (*i*) is similar in essential characteristics to its parent population; and (*ii*) reflects the salient features of the process by which it is generated. This heuristic is explicated in a series of empirical examples demonstrating predictable and systematic errors in the evaluation of uncertain events. In particular, since sample size does not represent any property of the population, it is expected to have little or no effect on judgment of likelihood. This prediction is confirmed in studies showing that subjective sampling distributions and posterior probability judgments are determined by the most salient characteristic of the sample (e.g., proportion, mean) without regard to the size of the sample. The present heuristic approach is contrasted with the normative (Bayesian) approach to the analysis of the judgment of uncertainty.

Subjective probabilities play an important role in our lives. The decisions we make, the conclusions we reach, and the explanations we offer are usually based on our judgments of the likelihood of uncertain events such as success in a new job, the outcome of an election, or the state of the market. Indeed an extensive experimental literature has been devoted to the question of how people perceive, process, and evaluate the probabilities of uncertain events in the contexts of probability learning, intuitive statistics, and decision making under risk. Although no systematic theory about the psychology of uncertainty has emerged from this literature, several empirical generalizations have been established. Perhaps the most general conclusion, obtained from numerous

---

430

investigations, is that people do not follow the principles of probability theory in judging the likelihood of uncertain events. This conclusion is hardly surprising because many of the laws of chance are neither intuitively apparent, nor easy to apply. Less obvious, however, is the fact that the deviations of subjective from objective probability[2] seem reliable, systematic, and difficult to eliminate. Apparently, people replace the laws of chance by heuristics, which sometimes yield reasonable estimates and quite often do not.

In the present paper, we investigate in detail one such heuristic called representativeness. A person who follows this heuristic evaluates the probability of an uncertain event, or a sample, by the degree to which it is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated. Our thesis is that, in many situations, an event A is judged more probable than an event B whenever A appears more representative than B. In other words, the ordering of events by their subjective probabilities coincides with their ordering by representativeness.

Representativeness, like perceptual similarity, is easier to assess than to characterize. In both cases, no general definition is available, yet there are many situations where people agree which of two stimuli is more similar to a standard, or which of two events is more representative of a given process. In this paper we do not scale representativeness, although this is a feasible approach. Instead, we consider cases where the ordering of events according to representativeness appears obvious, and show that people consistently judge the more representative event to be the more likely, whether it is or not. Although representativeness may play an important role in many varieties of probability judgments, e.g., political forecasting and clinical judgment, the present treatment is restricted to essentially repetitive situations where objective probabilities are readily computable.

Most data reported in this paper were collected in questionnaire form from a total of approximately 1500 respondents in Israel. The respondents were students in grades 10, 11, and 12 of college-preparatory high schools (ages 15–18). Special efforts were made to maintain the attention and the motivation of the Ss. The questionnaires were admin-

[2] We use the term "subjective probability" to denote any estimate of the probability of an event, which is given by a subject, or inferred from his behavior. These estimates are not assumed to satisfy any axioms or consistency requirements. We use the term "objective probability" to denote values calculated, on the basis of stated assumptions, according to the laws of the probability calculus. It should be evident that this terminology is noncommittal with respect to any philosophical view of probability.

istered in quiz-like fashion in a natural classroom situation, and the respondents' names were recorded on the answer sheets. Each respondent answered a small number (typically 2–4) of questions each of which required, at most, 2 min. The questions were introduced as a study of people's intuitions about chance. They were preceded by standard oral instructions which explained the appropriate question in detail. The experimental design was counterbalanced to prevent confounding with school or age. Most questions were pretested on University undergraduates (ages 20–25) and the results of the two populations were indistinguishable.

DETERMINANTS OF REPRESENTATIVENESS

In this section we discuss the characteristics of samples, or events, that make them representative, and demonstrate their effects on subjective probability. First, we describe some of the features that determine the similarity of a sample to its parent population. Then, we turn to the analysis of the determinants of apparent randomness.

Similarity of Sample to Population

The notion of representativeness is best explicated by specific examples. Consider the following question:

All families of six children in a city were surveyed. In 72 families the exact order *of births of boys and girls was* G B G B B G.
What is your estimate of the number of families surveyed in which the exact order *of births was* B G B B B B?

The two birth sequences are about equally likely, but most people will surely agree that they are not equally representative. The sequence with five boys and one girl fails to reflect the proportion of boys and girls in the population. Indeed, 75 of 92 Ss judged this sequence to be less likely than the standard sequence ($p < .01$ by a sign test). The median estimate was 30. Similar results have been reported by Cohen and Hansel (1956), and by Alberoni (1962).

One may wonder whether Ss do not simply ignore order information, and answer the question by evaluating the frequency of families of five boys and one girl relative to that of families of three boys and three girls. However, when we asked the same Ss to estimate the frequency of the sequence B B B G G G, they viewed it as significantly less likely than G B B G B G ($p < .01$), presumably because the former appears less random. Order information, therefore, is not simply ignored.

A related determinant of representativeness is whether the sample preserves the majority–minority relation in the population. We expect

a sample that preserves this relation to be judged more probable than an (objectively) equally likely sample where this relation is violated. This effect is illustrated in the following problem:

There are two programs in a high school. Boys are a majority (65%) in program A, and a minority (45%) in program B. There is an equal number of classes in each of the two programs.
You enter a class at random, and observe that 55% of the students are boys. What is your best guess—does the class belong to program A or to program B?

Since the majority of students in the class are boys, the class is more representative of program A than of program B. Accordingly, 67 of 89 Ss guessed that the class belongs to program A ($p < .01$ by sign test). In fact, it is slightly more likely that the class belongs to program B (since the variance for $p = .45$ exceeds that for $p = .65$).

A sample in which the various possible outcomes are present is, in general, more representative than a comparable sample in which some of the outcomes are not included. For example, given a binomial process with $p = \frac{4}{5}$, a significant majority of Ss judge a sample of 10 successes and 0 failures to be less likely than a sample of 6 successes and 4 failures, although the former sample is, in fact, more likely(see Figs. 1a and 2b).

The biasing effects of representativeness are not limited to naive subjects. They are also found (Tversky & Kahneman, 1971) in the intuitive judgments of sophisticated psychologists. Statistical significance is commonly viewed as the representation of scientific truth. Hence, a real effect (in the population) is expected to be represented by a significant result (in the sample) with insufficient regard for the size of the sample. As a consequence, researchers are prone to overestimate the likelihood of a significant result whenever they believe the null hypothesis to be false.

For example, the following question was posed to the participants of a meeting of the Mathematical Psychology Group and of the American Psychological Association:

Suppose you have run an experiment on 20 Ss, and have obtained a significant result which confirms your theory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 Ss. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?

A realistic estimate of the desired probability is somewhat lower than .50. The median estimate of the respondents was as high as .85. This unjustified confidence in the replicability of significance has severe con-

sequences for the conduct of research: it leads to unrealistic expectations concerning significance, and results in the planning of studies which are deficient in statistical power, see Cohen (1962).

Reflection of Randomness

To be representative, it is not sufficient that an uncertain event be similar to its parent population. The event should also reflect the properties of the uncertain process by which it is generated, that is, it should appear random. As is true of the similarity of sample to population, the specific features that determine apparent randomness differ depending on context. Nevertheless, two general properties, irregularity and local representativeness, seem to capture the intuitive notion of randomness. These properties are now discussed in turn.

A major characteristic of apparent randomness is the absence of systematic patterns. A sequence of coin[3] tosses, for example, which contains an obvious regularity is not representative. Thus, alternating sequences of heads and tails, such as H T H T H T H T or T T H H T T H H, fail to reflect the randomness of the process. Indeed, Ss judge such sequences as relatively unlikely and avoid them in producing simulated random sequences (Tune, 1964; Wagenaar, 1970).

Some irregularity is expected, not only in the order of outcomes, but also in their distribution, as shown in the following problem:

On each round of a game, 20 marbles are distributed at random among five children: Alan, Ben, Carl, Dan, and Ed. Consider the following distributions:

	I		II
Alan	4	Alan	4
Ben	4	Ben	4
Carl	5	Carl	4
Dan	4	Dan	4
Ed	3	Ed	4

In many rounds of the game, will there be more results of type I or of type II?

The uniform distribution of marbles (II) is, objectively, more probable than the nonuniform distribution (I), yet it appears too lawful to be the result of a random process. Distribution I, which departs slightly from an equitable partition, is more representative of random allocation. A significant majority of Ss (36 of 52, $p < .01$ by a sign test) viewed distribution I as more probable than distribution II. The presence of

---

[3] In this paper we deal with fair coins only.

some perturbation contributes to the representativeness and hence to the apparent likelihood of uncertain events.

Ss answer the above problem as if they ignored the individual nature of the two distributions and compared, instead, the two respective classes of distributions, disregarding the particular assignment of marbles to children. This does not mean that Ss do not appreciate the distinction between a class and its instances. What they do not appreciate is the proper impact of this distinction on judgments of relative frequency.

People view chance as unpredictable but essentially fair. Thus, they expect that in a purely random allocation of marbles each child will get approximately (though not exactly) the same number of marbles. Similarly, they expect even short sequences of coin tosses to include about the same number of heads and tails. More generally, a representative sample is one in which the essential characteristics of the parent population are represented not only globally in the entire sample, but also locally in each of its parts. A sample that is locally representative, however, deviates systematically from chance expectations: it contains too many alternations and too few clusters.

The law of large numbers ensures that very large samples are highly representative of the populations from which they are drawn. Elsewhere (Tversky & Kahneman, 1971), we have characterized the expectancy of local representativeness as a belief in the law of small numbers, according to which "the law of large numbers applies to small numbers as well." This belief, we suggest, underlies the erroneous intuitions about randomness, which are manifest in a wide variety of contexts.

Research on the perception of randomness (e.g., Tune, 1964; Wagenaar, 1970) shows that when people are asked to simulate a random process, such as a series of coin tosses, they produce sequences which are locally representative, with far too many short runs. Moreover, people tend to regard as unlikely, or reject as nonrandom, sequences which have the correct distribution of run lengths, presumably because long runs are not locally representative.

Similar findings have also been obtained in the hundreds of studies on probability learning and binary prediction (Estes, 1964; Jones, 1971). The gambler's fallacy, or the negative-recency effect, is a manifestation of the belief in local representativeness. For if the proportions of the two outcomes are to be preserved in short segments, then a long sequence of one outcome must be followed by the other outcome in order to restore the balance. In a locally representative world, in fact, the gambler's fallacy is no longer fallacious.

In his *Introduction to Probability Theory*, Feller (1968, p. 160) describes an example which illustrates the erroneous belief in local representativeness. During the intensive bombing of London in the Second

World War, it was generally believed that the bombing pattern could not be random, because a few sections of town were hit several times while many others were not hit at all. Thus, the pattern of hits violated local representativeness, and the randomness hypothesis seemed unacceptable. To test this hypothesis, the entire area of South London was divided into small sections of equal area, and the actual distribution of hits per section was compared to the expected (Poisson) distribution under the assumption of random bombing. Contrary to the general belief, the correspondence between the distributions was remarkably good. "To the untrained eye," Feller remarks, "randomness appears as regularity or tendency to cluster."

Most students are surprised to learn that in a group of as few as 23 people, the probability that at least two of them have the same birthday (i.e., same day and month) exceeds .5. Clearly, with 23 people the expected number of birthdays per day is less than $\frac{1}{15}$. Thus a day with two birthdays, in the presence of 343 "empty" days, is highly nonrepresentative, and the event in question, therefore, appears unlikely. More generally, we conjecture that the counterintuitive nature of many results in probability theory is attributable to violations of representativeness. (For a striking example from the theory of random walks, see Feller, 1968, pp. 84–88.)

A representative sample, then, is similar to the population in essential characteristics, and reflects randomness as people see it; that is, all its parts are representative and none is too regular. Only a few of all possible samples meet all these constraints. Most samples do not, and therefore do not appear random. Among the 20 possible sequences (disregarding direction and label) of six tosses of a coin, for example, we venture that only H T T H T H appears really random. For four tosses, there may not be any.

The tendency to regard some binary sequences as more random than others had dramatic consequences in the Zenith radio experiments[4] in which the audience was challenged to send in guesses of the identity of five binary symbols that were "telepathed" by a panel. The analysis of over a million responses (Goodfellow, 1938) revealed that the number of hits was far in excess of chance for some sequences and far below chance for others, depending largely on the apparent randomness of the target sequences. The implications of this finding for ESP research are obvious.

Random-appearing sequences are those whose verbal description is longest. Imagine yourself dictating a long sequence of binary symbols, say heads and tails. You will undoubtedly use shortcut expressions such

as "four Ts," or "H–T, three times." A sequence with many long runs allows shortcuts of the first type. A sequence with numerous short runs calls for shortcuts of the second type. The run structure of a random-appearing sequence minimizes the availability of these shortcuts, and hence defies economical descriptions. Apparent randomness, therefore, is a form of complexity of structure. Determinants of structural complexity, such as codability (Garner, 1970; Glanzer & Clark, 1963; Vitz & Todd, 1969) affect apparent randomness as well.

SAMPLING DISTRIBUTIONS

We have proposed that Ss assign probabilities to events so that the more representative events are assigned higher probabilities, and equally representative events are assigned equal probabilities. In this section, we investigate the implication of this hypothesis for the study of subjective sampling distributions, i.e., the probabilities that Ss assign to samples of a given size from a specified population.

When the sample is described in terms of a single statistic, e.g., proportion or mean, the degree to which it represents the population is determined by the similarity of that statistic to the corresponding parameter of the population. Since the size of the sample does not reflect any property of the parent population, it does not affect representativeness. Thus, the event of finding more than 600 boys in a sample of 1000 babies, for example, is as representative as the event of finding more than 60 boys in a sample of 100 babies. The two events, therefore, would be judged equally probable, although the latter, in fact, is vastly more likely. Similarly, according to the present analysis, the subjective probabilities that the average height in a sample of men lies between 6 ft 0 in. and 6 ft 2 in. would be independent of the size of the sample.

To test these predictions, nine different groups of Ss produced subjective sampling distributions for three sample sizes ($N = 10, 100, 1000$) and for each of the following three populations.

Distribution of sexes. (Binomial, $p = .50$) Ss were told that approximately N babies are born every day in a certain region. For $N = 1000$, for instance, the question read as follows:

> On what percentage of days will the number of boys among 1000 babies be as follows:
>   Up to 50 boys
>   50 to 150 boys
>   150 to 250 boys
>   . . . .
>   850 to 950 boys
>   More than 950 boys
> Note that the categories include all possibilities, so your answers should add up to about 100%.

For $N = 100$, the 11 categories were: up to 5, 5–15, etc. For $N = 10$, each category contained a single outcome, e.g., 6 boys.

Distribution of heartbeat type. (Binomial, $p = .80$) Here, Ss were told that approximately N babies are born every day in a certain region, and that 80% of all newborns have a heartbeat of type α and the remaining 20% have a heartbeat of type β. For each sample size, Ss produced sampling distributions for the number of babies born every day with heartbeat of type α using the same 11 categories as above.

Distribution of height. Ss were told that a regional induction center records the average height of the N men who are examined every day. They were also told that the average height of the male population lies between 170–175 cm (in Israel height is measured in centimeters), and that the frequency of heights decreases with the distance from the mean. For each sample size, Ss produced a sampling distribution of average height, in the following seven categories: up to 160, 160–165, . . . , more than 185.
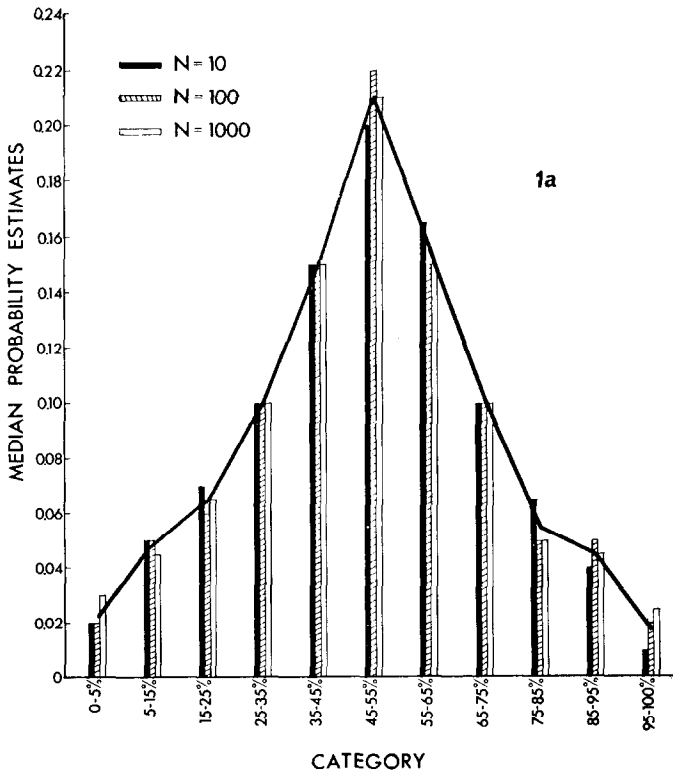

FIGURE 1a

Median estimates for the three populations, respectively, are shown in Figs. 1a, b, and c, for all three values of N. (Size of group varied from 45 to 84 with an average of 62.) It is apparent that sample size has no effect whatsoever on the subjective sampling distributions. Independent groups, faced with problems that differ only in sample size, produce indistinguishable distributions. This result holds for populations that are defined abstractly, e.g., the binomial, as well as for populations that are known to Ss through daily experience, e.g., the height of men.

Since subjective sampling distributions are independent of N, the solid lines in each figure, which connect the means of the median estimates, can be regarded as "universal" sampling distributions for the respective population. To depict the magnitude of the true effect of sample size, which Ss completely ignore, the correct sampling distributions for $p = .50$ and $p = .80$ are shown, together with the corresponding "universal" sampling distribution, in Figs. 2a and b, respectively.

It can be seen that the "universal" curves are even flatter than the correct curves for $N = 10$. For $p = .50$, the "universal" variance (.048) is roughly equal to the correct sampling variance for $N = 5$ (.05). For
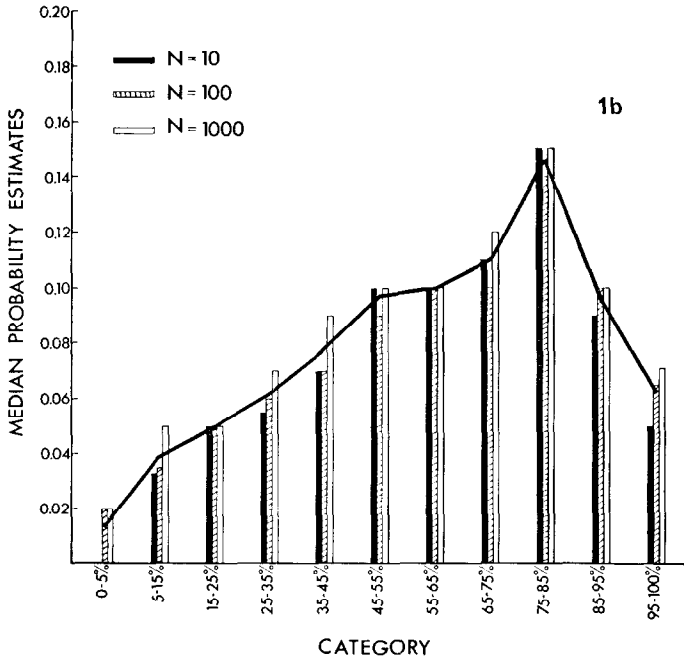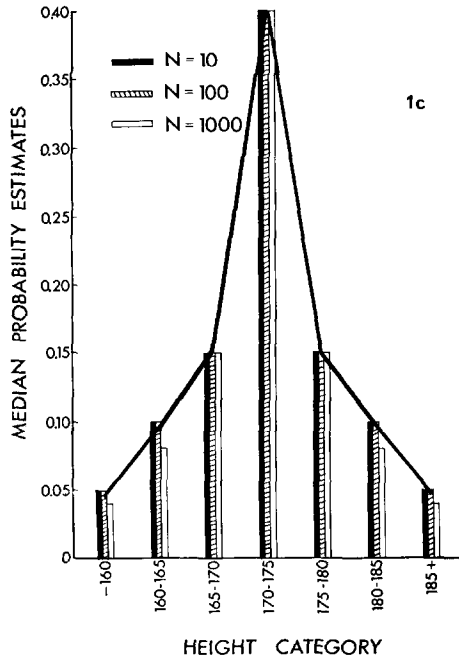

FIGURE 1b

HEIGHT CATEGORY

FIGURE 1c

$p = .80$, the variance of the "universal" curve (.068) lies between the correct sampling variance for $N = 2$ and that for $N = 3$.

In binomial distributions, the mean generally coincides with the mode. Consequently, when $p \neq .50$, the short tail must be higher than the long tail; see, for example, the correct distribution for $N = 10$ in Fig. 2b. Figure 2b also shows that this property is violated by the "universal" curve for $p = .80$ whose mean is only .63. Thus, although the mode of the subjective sampling distribution is properly located at the most representative value, the mean is displaced towards the long tail. The same result has been obtained in other studies, e.g., Cohen and Hansel (1956), Peterson, DuCharme, and Edwards (1968). Thus, for $p = .80$ the "universal" sampling distribution of the proportion is not a binomial at all!

The present experiment differs from previous studies of the subjective binomial (Peterson, DuCharme, & Edwards, 1968; Wheeler & Beach, 1968) in two respects. First, the earlier work was concerned with sample sizes much smaller than those of the present study. Second, and more important, the number of events among which probabilities were distributed was not the same for different sample sizes: for a
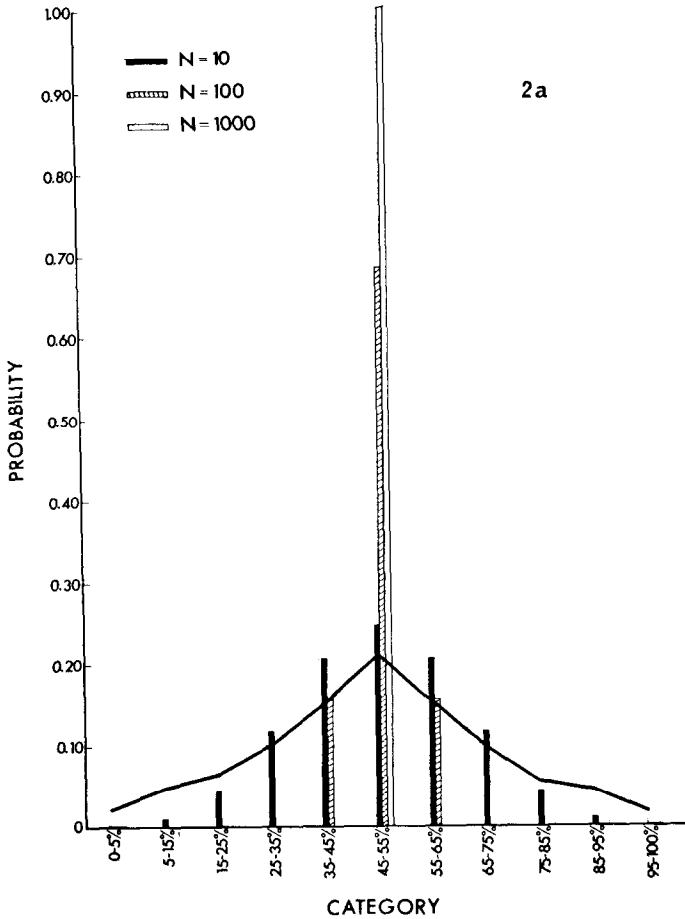
FIGURE 2a

sample of size N, Ss evaluated $N + 1$ outcomes. In the present study, in contrast, Ss evaluate the *same* number of categories for all sample sizes. The invariance of the subjective sampling distribution with respect to N, which is demonstrated in Fig. 1, may not hold exactly when the number of categories varies, or when the sample is small enough to permit enumeration of possibilities. For larger samples, enumeration is impossible, and the natural recourse is to a direct appreciation of representativeness, which is dominated by sample mean, or sample proportion.

To further explore the representativeness prediction concerning sample size, an additional experiment was conducted. Ss were 97 Stanford
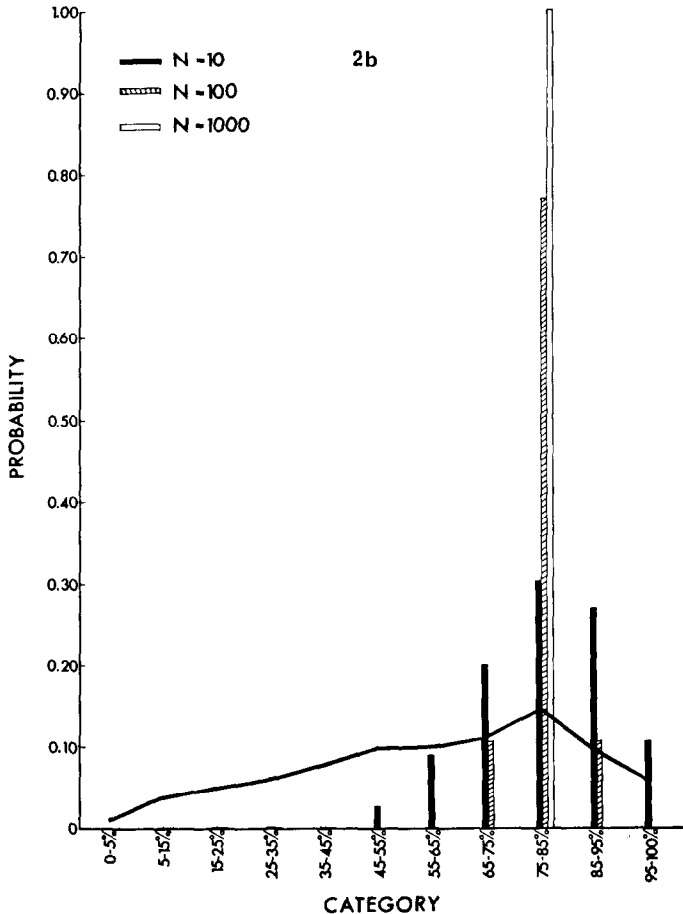
FIGURE 2b

undergraduates with no background in probability or statistics, run in small groups of 5 to 12 members each. Ss were presented, in a fixed order, with three problems each defining a sampling process with a specified mean and a critical value above that mean, and asked to judge whether a particular sampling outcome is more likely to occur in a small or in a large sample. Each S was paid $1 for participation in the experiment and an additional $1 if his answer to one of the problems (randomly selected after completion of the task) was correct.

To control for response bias, each problem was presented in two forms. Half the Ss judged, for all three problems, whether an outcome that is *more* extreme than the specified critical value is more likely to

occur in a small or in a large sample. The correct answer, of course, is that an extreme outcome is more likely to occur in a small sample. The remaining Ss judged whether an outcome that is *less* extreme than the specified critical value is more likely to occur in a small or in a large sample. The correct answer here is that such an outcome is more likely to occur in a large sample. The three problems are presented below. The values in parentheses are the numbers of Ss who chose each response category, for each of the two forms. The correct answers are starred.

1. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of 1 year, each hospital recorded the days on which (more/less) than 60% of the babies born were boys. Which hospital do you think recorded more such days?

	(More than 60%)	(Less than 60%)
The larger hospital	(12)	(9)*
The smaller hospital	(10)*	(11)
About the same (i.e., within 5% of each other)	(28)	(25)

2. An investigator studying some properties of language selected a paperback and computed the average word-length in every page of the book (i.e., the number of letters in that page divided by the number of words). Another investigator took the first line in each page and computed the line's average word-length. The average word-length in the entire book is four. However, not every line or page has exactly that average. Some may have a higher average word-length, some lower.

The first investigator counted the number of pages that had an average word-length of 6 or (more/less) and the second investigator counted the number of lines that had an average word-length of 6 or (more/less). Which investigator do you think recorded a larger number of such units (pages for one, lines for the other)?

	(More than 6)	(Less than 6)
The page investigator	(8)	(10)*
The line investigator	(21)*	(15)
About the same (i.e., within 5% of each other)	(20)	(23)

3. A medical survey is being held to study some factors pertaining to coronary diseases. Two teams are collecting data. One checks three men a day, and the other checks one man a day. These men are chosen randomly from the population. Each man's height is measured during the checkup. The average height of adult males is 5 ft 10 in., and there are as many men whose height is above average as there are men whose height is below average.

The team checking three men a day ranks them with respect to their height, and counts the days on which the height of the middle man is (more/less) than 5 ft 11 in. The other team merely counts the days on which the man they checked was (taller/shorter) than 5 ft 11 in. Which team do you think counted more such days?

	(More than 5 ft 11 in.)	(Less than 5 ft 11 in.)
The team checking 3	(7)	(14)*
The team checking 1	(18)*	(17)
About the same (i.e., within 5% of each other)	(23)	(17)

If Ss have any insight into the role of sample size, they should find it easy to select the correct answers to these simple ordinal questions. On the other hand, if they judge equally representative outcomes to be equally likely, they should show no systematic preference for the correct answer. This is clearly the case. The modal answer is "same" in almost all comparisons; moreover, there is no significant preference for the correct answer in any of the problems.

This experiment confirms the conclusions of the initial study in spite of several procedural differences. Here, each S makes a direct ordinal judgment of the likelihood of an outcome with two sample sizes under conditions designed to motivate accuracy. This procedure should enhance the salience of sample size. Furthermore, the last problem compares a single observation to the median of a sample of three observations. Apparently, Ss fail to notice even the obvious fact that medians must be less variable than single observations.

The notion that sampling variance decreases in proportion to sample size is apparently not part of man's repertoire of intuitions. Indeed, misconceptions of the role of sample size occur frequently in everyday life. On the one hand, people are often willing to take seriously a result stated in percentages, with no concern for the number of observations, which may be ridiculously small. On the other hand, people often remain skeptical in the face of solid evidence from a large sample, as in the case of the well-known politician who complained bitterly that the cost-of-living index is not based on the whole population, but only on a large sample, and added, "Worse yet—a *random* sample."

We surely do not mean to imply that man is incapable of appreciating the impact of sample size on sampling variance. People can be taught the correct rule, perhaps even with little difficulty. The point remains that people do not follow the correct rule, when left to their own devices.[5] Furthermore, the study of the conduct of research psy-

---

[5] In this respect, the present sampling task may be similar to a deductive reasoning task investigated by Johnson-Laird and Wason (1970) in which Ss are incapable of discovering a very simple logical rule which they readily understand and apply once it is called to their attention.

chologists (Cohen, 1962; Tversky & Kahneman, 1971) reveals that a strong tendency to underestimate the impact of sample size lingers on despite knowledge of the correct rule and extensive statistical training. For anyone who would wish to view man as a reasonable intuitive statistician, such results are discouraging.

POSTERIOR PROBABILITY

In the problems studied in the previous section, Ss estimated the probability that a certain sample had been drawn from a given population. A related task, which has become increasingly popular in recent years under the impact of the Bayesian approach to statistical inference, is the evaluation of *posterior* probability. In this task, Ss estimate the probability, or the odds, that a given sample has been drawn from one rather than another population. A typical experiment is run as follows: the S is shown two bookbags, one containing, say, 80% red poker chips and 20% blue poker chips, and the other containing reversed proportions of red and blue chips. One of the bags is selected by chance and a random sample is drawn from it. S observes the number of red and blue chips in the sample, and estimates the posterior probability, or the odds, that the sample has been drawn from the predominantly red bag. Alternatively, the S may be required to evaluate the posterior probability that a sample of observations, say measurements of height, has been drawn from a population of men rather than of women.

To obtain the correct answer, one computes the probabilities of obtaining the observed sample under each of the two hypotheses. The objective posterior odds in such problems are simply the ratio of these two probabilities, known as the likelihood ratio. The performance of Ss as intuitive estimators of posterior probability has been the subject of an extensive experimental literature, reviewed by Peterson and Beach (1967), Edwards (1968), Slovic and Lichtenstein (1971), and Rapoport and Wallsten (1972).

Posterior probabilities have been investigated in two different contexts: the revision of opinion and the evaluation of evidence. Revision of opinion is studied with a sequential procedure in which S is presented with successive data and revises his posterior estimate after each datum. Evaluation of evidence is studied with an aggregate procedure in which S is presented with the entire data at once and makes a single estimate of posterior probability.

The most obvious fact about the sequential procedure is that Ss generally revise their opinions in the correct direction after each datum: a red chip increases their confidence that the sample has been drawn from the predominantly red bag. Consequently, the subjective posterior estimates appear monotonically related to the objective posterior prob-

abilities. The subjective odds, however, are generally conservative, i.e., too close to unity.[6] These findings have fostered an approach which adopts the normative Bayesian rule as a basic model of the behavior of the S. Man is viewed as a conservative Bayesian estimator, and his deviations from the norm are attributed to misperception of the impact of each datum, misaggregation of the joint impact of data, or to a response bias against extreme estimates (Edwards, 1968). All three hypotheses predict that, within any experimental situation, subjective posterior estimates should be monotonically related to the correct Bayesian values. According to this approach, therefore, Ss' estimates are assumed to be qualitatively compatible with the normative model: they merely exhibit a conservative bias that is to be explained.

The invariance of subjective sampling distributions with respect to sample size has led us to question the adequacy of the normative approach to the modeling of intuitive statistical judgments. The analysis of subjective sampling distributions in terms of representativeness extends naturally to subjective posterior probabilities in the context of the evaluation of evidence. (Since the revision of opinion in the sequential procedure introduces additional complications, we focus primarily on the simpler aggregate procedure.) Earlier we argued that the subjective probability that a certain sample has been drawn from a given population is determined by the representativeness of the most salient feature of that sample. Extending the analysis of sampling distributions to posterior probability judgment, we propose that the subjective probability that a sample has been drawn from one rather than another population is a function of the degree to which the sample is representative of each of the populations. It follows from this hypothesis (with no additional assumptions about the form of the function) that subjective posterior probability should depend on the *same* salient characteristics of the sample which determine subjective sampling distributions.

Consider the binomial problem, for example. The most salient property of a binomial sample is, clearly, the sample proportion. Hence we expect, at least in the aggregate case, that the subjective posterior probability will depend primarily on the proportion of red chips, say, in the sample with little or no regard to the size of the sample or to the difference between the proportions of red chips in the two symmetric populations.

The objective situation, of course, is quite different. In the symmetric binomial task the objective posterior probability depends only on the

[6] Damon Runyon tells of a gambler who claims that no proposition is better than 8 to 5. His betting, however, could hardly be described as conservative.

difference between the numbers of red and blue chips observed in the sample. Specifically, the posterior odds are given by $(p/1-p)^{r-b}$, where p is the proportion of red chips in the population and r and b are, respectively, the numbers of red and blue chips observed in the sample. Thus, a sample of 5 red chips and 1 blue chip, and a sample of 15 red and 11 blue chips both give equal support to the hypothesis that the sample was drawn from the predominantly red bag. This result seems counterintuitive because the former sample, in which the proportion of red chips is $5/6$, seems to provide much stronger evidence for the hypothesis that the majority is red than the second sample, in which the proportion of red chips is only $15/26$.

To test the above prediction, we constructed a set of 10 symmetric binomial problems in which population proportion, sample ratio, and sample difference were varied systematically. A typical problem reads as follows:

Consider two very large decks of cards, denoted A and B. In deck A, $5/6$ of the cards are marked X, and $1/6$ are marked O. In deck B, $1/6$ of the cards are marked X, and $5/6$ are marked O.

One of the decks has been selected by chance, and 12 cards have been drawn at random from it, of which 8 are marked X and 4 are marked O.

What do you think the probability is that the 12 cards were drawn from deck A, that is, from the deck in which most of the cards are marked X?

In this problem, the sample ratio is 8:4 and the sample difference is 4. Two other problems maintained the same sample ratio: 4:2 and 40:20, and two problems maintained the same difference: 5:1 and 18:14. Five additional problems were obtained by replacing the population proportions of $5/6$ and $1/6$, respectively, by $2/3$ and $1/3$.

Each of the 10 problems was presented to a different group of Ss. Size of group varied from 37 to 79 with an average of 56. Ss received a general explanation concerning the task, and it was pointed out to them that if a sample contains more Xs than Os, then the probability that it was drawn from deck A is at least .50. Nevertheless, about 10% of the Ss violated this rule, and were excluded from the analysis.

Table 1 presents the results separately for the two pairs of populations. In each case, the three problems in the middle row have the same sample difference, and the three problems in the middle column have the same sample ratio. Understandably perhaps, Ss found the question rather difficult, and most restricted their answers to multiples of .10. Nevertheless, the pattern of results is perfectly clear. First, population proportions (p) have no effect whatsoever on the subjective posterior

TABLE 1

Subjective Posterior Probability in a Symmetric Binomial Task for Two Pairs of Populations: $p = 5/6$ and $p = 2/3$

The upper entry in each cell is the sample presented; the lower entry is the median subjective estimate.

	$p = 5/6$			$p = 2/3$		
		4:2 .70			4:2 .68	
Sample Difference	18:14 .60	8:4 .70	5:1 .83	18:14 .58	8:4 .70	5:1 .85
		40:20 .70			40:20 .70	

Sample ratio

estimates. (From a normative standpoint, of course, p should have a substantial effect. The posterior odds for the 4:2 sample, for instance, are 4 when $p = 2/3$, and 25 when $p = 5/6$).

Second, it is evident that the observed estimates depend on sample ratio and do not depend, as they should objectively, on sample difference. For both p levels, the estimates for the ratio 5:1 are significantly higher than those in the middle column (where the ratio is 2:1) and these in turn are significantly higher than the estimates for the ratio 18:14. (In all comparisons, $p < .01$, by median test.) In fact, median posterior estimates are very close to the stated sample proportion in all 10 problems.

S's estimates violate the Bayesian rule qualitatively as well as quantitatively. Not only are the estimates systematically conservative (all but one objective probability exceed .94), but they are actually determined by the wrong variable. Consequently, the subjective and objective estimates are not even monotonically related. The 40:20 sample, for example, provides considerably stronger objective evidence than the 5:1 sample, but Ss nonetheless view the latter as much more convincing.

The finding that, in the symmetric binomial task, subjective posterior estimates depend on the content of the sample but not on the population proportion has been obtained in several studies using both aggregate and sequential procedures (Vlek, 1965; Pitz, Downing, & Reinhold, 1967; Beach, Wise, & Barclay, 1970). The finding that, in the

aggregate case, subjective posterior estimates are determined by sample ratio rather than by sample difference has also been reported in several recent studies (Shanteau, 1970; Beach, Wise, & Barclay, 1970). In the sequential procedure, however, both sample ratio and sample difference affect the posterior estimates (Vlek, 1965; Peterson, Schneider, & Miller, 1965; Pitz, 1967; Beach, Wise, & Barclay, 1970). The effect of sample difference in the sequential case may be due to the fact that consecutive drawings from a given bag yield successive samples in which the ratio remains stochastically constant while the difference increases. The increase of subjective probability with increasing sample difference, in this case, is attributable to Ss' tendency to revise their estimates by a constant amount according to the last item of information (Pitz, Downing, & Reinhold, 1967; Shanteau, 1970).

The tendency to evaluate posterior probability in terms of the most salient features of the sample, with little regard to additional factors, is not limited to binomial problems. The following problem was posed to 115 students from the University of Michigan, all of whom had had at least one course in statistics.

The average heights of adult males and females in the US are, respectively, 5 ft 10 in. and 5 ft 4 in. Both distributions are approximately normal with a standard deviation of about 2.5 in.

An investigator has selected one population by chance and has drawn from it a random sample.

What do you think are the odds that he has selected the male population if
(i) the sample consists of a single person whose height is 5 ft 10 in.?
(ii) the sample consists of 6 persons whose average height is 5 ft 8 in.?

The median subjective odds were 8 in case (i) and 2.5 in case (ii). Indeed, the significant majority of Ss (86 out of 115) assigned a higher value to the former case ($p < .01$ by a median test). The correct odds are 16 in case (i) and 29 in case (ii). The responses of the Ss, therefore, are not merely conservative—they violate the correct ordering of likelihoods. Here again, it appears that Ss base their judgments on sample mean with insufficient concern for sample size.

NORMATIVE MODELS AND DESCRIPTIVE HEURISTICS

The view has been expressed (see, e.g., Edwards, 1968) that man, by and large, follows the correct Bayesian rule, but fails to appreciate the full impact of evidence, and is therefore conservative. Peterson and Beach (1967), for example, concluded that the normative model provides a good first approximation to the behavior of the Ss who are "influenced by appropriate variables and in appropriate directions [p.

43]." This view has not been shared by all. In a more recent review of the literature, Slovic and Lichtenstein (1971) argued that the above evaluation of man's performance as an intuitive statistician is far "too generous," while Pitz, Downing, and Reinhold (1967) concluded, on the basis of their data, that human performance in Bayesian tasks is "nonoptimal in a more fundamental way than is implied by discussions of conservatism [p. 392]."

The usefulness of the normative Bayesian approach to the analysis and the modeling of subjective probability depends primarily not on the accuracy of the subjective estimates, but rather on whether the model captures the essential determinants of the judgment process. The research discussed in this paper suggests that it does not. In particular, we have seen that sample size has no effect on subjective sampling distributions, that posterior binomial estimates are determined (in the aggregate case, at least) by sample proportion rather than by sample difference, and that they do not depend on the population proportion. In his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all.

It could be argued that the failure of the normative model to describe human behavior is limited to naive Ss faced wth unfamiliar random processes, and that the normative model could provide an adequate account of the evaluation of the more familiar random processes that people encounter in everyday life. There is very little evidence, however, to support this view. First, it has been shown (Tversky & Kahnemann, 1971) that the same type of systematic errors that are suggested by considerations of representativeness can be found in the intuitive judgments of sophisticated scientists. Apparently, acquaintance with the theory of probability does not eliminate all erroneous intuitions concerning the laws of chance. Second, in our daily life we encounter numerous random processes (e.g., the birth of a boy or a girl, hitting a red light at a given intersection, getting a hand with no hearts in a card game) which obey the binomial law, for example, to a high degree of approximation. People, however, fail to extract from these experiences an adequate conception of the binomial process. Apparently, extensive exposure to numerous examples alone does not produce optimal behavior.

In their daily lives, people ask themselves and others questions such as: What are the chances that this 12-year-old boy will grow up to be a scientist? What is the probability that this candidate will be elected to office? What is the likelihood that this company will go out of business? These problems differ from those discussed earlier in the paper in that, due to their unique character, they cannot be readily answered

either in terms of frequency of occurrence in the past, or in terms of some well-defined sampling process.

In this paper, we investigated in some detail one heuristic according to which the likelihood of an event is evaluated by the degree to which it is representative of the major characteristics of the process or population from which it originated. Although our experimental examples were confined to well-defined sampling processes (where objective probability is readily computable), we conjecture that the same heuristic plays an important role in the evaluation of uncertainty in essentially unique situations where no "correct" answer is available. The likelihood that a particular 12-year-old boy will become a scientist, for example, may be evaluated by the degree to which the role of a scientist is representative of our image of the boy. Similarly, in thinking about the chances that a company will go out of business, or that a politician will be elected for office, we have in mind a model of the company, or of the political situation, and we evaluate as most likely those outcomes which best represent the essential features of the corresponding model.

The assessment of representativeness is not the only way to establish an intuitive judgment of probability. Elsewhere (Tversky & Kahneman, 1972), we have investigated another heuristic—availability—according to which one judges the probability of an event (e.g., snow in November) by the ease with which relevant instances (e.g., past Novembers) are recalled or imagined. More specifically, we propose that the number of relevant instances that could be readily retrieved or the ease with which they come to mind are major clues that men use in estimating probability or frequency. For example, suppose you sample a word (containing three or more letters) from an English text. Is it more likely that the word starts with a "k," or that it has a "k" in the third position? To answer such a question, people often try to think of words beginning with "k" (e.g., key) and of words that have "k" in third position (e.g., like), and then compare the number or the ease with which the two types of words come to mind. Obviously, it is easier to think of words that start with a "k" than of words with a "k" in the third position. Indeed, the majority of Ss judged the former event more likely despite the fact that there are three times as many words with a "k" in the third position (Tversky & Kahneman, 1972).

The major difference between the two heuristics lies in the nature of the judgment which underlies the evaluation of subjective probability. According to the representativeness heuristic, one evaluates subjective probability by the degree of correspondence between the sample and the population, or between an occurrence and a model. This heuristic, therefore, emphasizes the generic features, or the connotation, of the

event. According to the availability heuristic, on the other hand, subjective probability is evaluated by the difficulty of retrieval and construction of instances. It focuses, therefore, on the particular instances, or the denotation, of the event. Thus, the representativeness heuristic is more likely to be employed when events are characterized in terms of their general properties; whereas the availability heuristic is more likely to be employed when events are more naturally thought of in terms of specific occurrences. When the generic features of an event as well as its specific instances are considered, both heuristics are likely to enter into the evaluation. In contemplating an event, such as the termination of a particular war by a given time, people often construct or imagine scenarios that lead to the event in question. The likelihood of the event may, then, be evaluated by the availability of scenarios that lead to ending the war as well as by the degree to which these scenarios are representative of the relevant military and political situation.

One general notion shared by both heuristics is the use of mental effort—of one form or another—to gauge subjective probability. It is certainly harder to imagine an uncertain process yielding a nonrepresentative outcome than to imagine the same process yielding a highly representative outcome. The former outcome, of course, produces more surprise, which is a form of mental effort. Similarly, the less available the instances of an event, the harder it is to retrieve and construct them, and the lower the judged probability of that event.

The results and speculations presented in this paper provide merely an outline of a heuristic approach to the study of man's competence and performance as a judge of uncertainty. The data base is admittedly narrow, and the more interesting problems of evaluating uncertainty in everyday life have yet to be faced. Nevertheless, it is evident that the present approach differs markedly from the normative approach in that it focuses on the question "how do people evaluate uncertainty?" rather than on "how *well* do people evaluate uncertainty?" We consider the former question to be of greater interest and importance particularly because, as was pointed out by Slovic and Lichtenstein (1971), optimal performance by a S is often attributable to "a fortuitous interaction between the optimal strategy and whatever simple rule he arrived at in his groping attempts to ease cognitive strain." Heuristics such as representativeness and availability, however, are not likely to be adopted *merely* because they reduce cognitive strain. They are probably adopted because they often work with a rate of success that is more than fortuitous. The study of the conditions under which the various heuristics work, and the manner in which they could be improved are promising areas for both theoretical and applied research.

REFERENCES

ALBERONI, F. Contribution to the study of subjective probability. Part I. *Journal of General Psychology,* 1962, **66,** 241–264.

BEACH, L. R., WISE, J. A., & BARCLAY, S. Sample proportion and subjective probability revisions. *Organizational Behavior and Human Performance,* 1970, **5,** 183–190.

COHEN, J. The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology,* 1962, **65,** 145–153.

COHEN, J., & HANSEL, C. E. M. *Risk and gambling.* New York: Philosophical Library, 1956.

EDWARDS, W. Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment.* New York: Wiley, 1968. Pp. 17–52.

ESTES, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press, 1964. Pp. 89–128.

FELLER, W. *An introduction to probability theory and its applications.* (3rd ed.) Vol. 1. New York: Wiley, 1968.

GARNER, W. R. Good patterns have few alternatives. *American Scientist,* 1970, **58,** 34–43.

GLANZER, M., & CLARK, W. H. Accuracy of perceptual recall: An analysis of organization. *Journal of Verbal Learning and Verbal Behavior,* 1963, **1,** 289–299.

GOODFELLOW, L. D. A psychological interpretation of the results of the Zenith radio experiments in telepathy. *Journal of Experimental Psychology,* 1938, **23,** 601–632.

JOHNSON-LAIRD, P. N., & WASON, P. C. A theoretical analysis of insight into a reasoning task. *Cognitive Psychology,* 1970, **1,** 134–148.

JONES, M. R. From probability learning to sequential processing: A critical review. *Psychological Bulletin,* 1971, **76,** 153–185.

PETERSON, C. R., & BEACH, L. R. Man as an intuitive statistician. *Psychological Bulletin,* 1967, **68,** 29–46.

PETERSON, C. R., DuCHARME, W. M., & EDWARDS, W. Sampling distributions and probability revisions. *Journal of Experimental Psychology,* 1968, **76,** 236–243.

PETERSON, C. R., SCHNEIDER, R. J., & MILLER, A. J. Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology,* 1965, **69,** 522–527.

PITZ, G. F. Sample size, likelihood, and confidence in a decision. *Psychonomic Science* 1967, **8,** 257–258.

PITZ, G. F., DOWNING, L., & REINHOLD, H. Sequential effects in the revision of subjective probabilities. *Canadian Journal of Psychology,* 1967, **21,** 381–393.

RAPOPORT, A., & WALLSTEN, T. S. Individual decision behavior. (Vol. 23) in *Annual Review of Psychology.* Palo Alto: Annual Reviews, 1972. Pp. 131–176.

SHANTEAU, J. C. An additive model for decision-making *Journal of Experimental Psychology,* 1970, **85,** 181–191.

SLOVIC, P., & LICHTENSTEIN, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance,* 1971, **6,** 649–744.

TUNE, G. S. Response preferences: A review of some relevant literature. *Psychological Bulletin,* 1964, **61,** 286–302.

TVERSKY, A., & KAHNEMAN, D. The belief in the law of small numbers. *Psychological Bulletin,* 1971, **76,** 105–110.

TVERSKY, A., & KAHNEMAN, D. Availability as a determinant of frequency and probability judgments. *Oregon Research Institute Technical Report* 1972, **12**(1).

VITZ, P. C., & TODD, T. C. A coded element model of the perceptual processing of sequential stimuli. *Psychological Review,* 1969, **76**, 433–449.

VLEK, C. The use of probabilistic information in decision making. Psychological Institute Report No. 009-65, University of Leiden, The Netherlands, 1965.

WAGENAAR, W. A. Subjective randomness and the capacity to generate information. In A. F. Sanders (Ed.) *Attention and performance III, Acta Psychologica,* 1970, **33**, 233–242.

WHEELER, G., & BEACH, L. R. Subjective sampling distributions and conservatism. *Organizational Behavior and Human Performance,* 1968, **3**, 36–46.