COGNITION AND INSTRUCTION, 1989, *6*(1), 59–98

Informal Conceptions of Probability

Clifford Konold
Scientific Reasoning Research Institute
University of Massachusetts

A model of informal reasoning under conditions of uncertainty, the *outcome approach,* was developed to account for the nonnormative responses of a subset of 16 undergraduates who were interviewed. For individuals who reason according to the outcome approach, the goal in questions of uncertainty is to predict the outcome of an individual trial. Their predictions take the form of yes–no decisions on whether an outcome will occur on a particular trial. These predictions are then evaluated as having been either right or wrong. Their predictions are often based on a deterministic model of the situation. In follow-up interviews using a different set of problems, responses of outcome-oriented participants were predicted. In one problem, their responses were at variance both with normative interpretations of probability and with the "representativeness heuristic" (Kahneman & Tversky, 1972). Although the outcome approach is inconsistent with formal theories of probability, its components are logically consistent and reasonable in the context of everyday decision making.

A weather forecast includes the prediction of a 70% chance of rain; lotteries and sweepstakes publish the odds of winning; studies report the failure rates of various methods of birth control. Information of this sort is intended to help people make more reasonable decisions. Yet recent research on human decision making in situations involving uncertainty has revealed that peoples' judgments are frequently not in agreement with probability and statistical theory (Kahneman, Slovic, & Tversky, 1982; Peterson & Beach, 1967; Pollatsek, Konold, Well, & Lima, 1984).

Amos Tversky and Daniel Kahneman have provided the most integrative account to date of the discrepancies between normative and actual judgments under uncertainty. According to Tversky and Kahneman (1983), two general types of cognitions are potentially available in making probabilistic

Requests for reprints should be sent to Clifford Konold, Scientific Reasoning Research Institute, Hasbrouck Laboratory, University of Massachusetts, Amherst, MA 01003.

judgments. On the one hand, people have acquired some knowledge of random events and basic probability theory that allows them to calculate the chance of various events in simple chance setups. Most people know, for example, that $p(A) + p(\overline{A}) = 1$, and that for setups with equally likely outcomes, the probability of a particular event is equal to the number of outcomes favorable to that event divided by the total number of equally likely outcomes (the classical interpretation of probability). Piaget and Inhelder (1951/1975) concluded that, by the age of 12, most children can reason probabilistically about a variety of random generating devices.

In addition to these capabilities, however, people have developed a number of judgment heuristics for analyzing complex, real-world events. These heuristics, according to Tversky and Kahneman (1983), are based on a collection of "natural assessments" that operate on a nonconscious, perceptual level. Although many decisions based on natural assessments are congruent with those that would be made on the basis of probability theory, there are many situations for which this is not true, where the perceptual processes and associated judgment heuristics lead to "statistical illusions." Kahneman and Tversky developed problems for which the use of these heuristics led to judgment errors, and used high rates of nonnormative responses to these problems to support the claim that, for those untrained in formal probability, these heuristics form the basis of probability judgments.

One example of a heuristic-based judgment error is the belief that the sequence MMMMMM of male and female births in a family is less likely than the sequence MFFMMF. Kahneman and Tversky (1972) suggested that this conclusion is reached through applying a "representativeness heuristic," according to which the probability of a sample is estimated by noting the degree of similarity between the sample and parent population. Because the sequence MFFMMF is more similar to the population proportion of approximately half males and half females and because it appears to better reflect the random process underlying sex determination, it is judged as more likely. According to probability theory, all possible sequences are equally likely.

Asked to compare the frequency of words in the English language that begin with *r* with those that have *r* as the third letter, most people incorrectly conclude that the former are more frequent. According to Tversky and Kahneman (1973), this judgment is made via the "availability heuristic," according to which the probability or frequency of an event is related to the ease or difficulty of recalling relevant instances of that event. Because it is easier for most people to search mentally for words according to their first letter, they mistakenly conclude that words beginning with *r* occur more frequently.

When people make probabilistic decisions, both the collection of natural assessments and more formal, conceptual knowledge of probability theory are presumably available. Which of these is applied in a particular instance is a function not only of individual differences in knowledge of probability theory but also of situation variables. Nisbett, Krantz, Jepson, and Kunda (1983) argued that people with little formal training in probability tend to analyze a situation probabilistically when (a) the sample space is easily recognizable (e.g., when the event is repeatable and outcomes are symmetric) and (b) the role of chance is salient (e.g., in coin flipping and urn drawing). On the other hand, even people who have had considerable training in the application of probabilistic models can be led to the unconscious application of natural assessments for situations that they know call for a probabilistic analysis (Tversky & Kahneman, 1971).

FORMAL AND INFORMAL CONCEPTIONS OF PROBABILITY

Hidden in the preceding account is the assumption that, regardless of whether one uses heuristics or formal probability knowledge, the individual has the goal of arriving at the probability of the event in question. Although the value that is finally arrived at may be nonnormative, the meaning of the value is assumed to lie somewhere in the range of acceptable interpretation.

The purpose of this study was to explore the possibility that errors in reasoning under uncertainty arise not only from indiscriminate application of natural assessments, but also from analyses based on a different understanding of the goal in reasoning under uncertainty. This hypothesis was formulated on the basis of observations made in an earlier study (Well, Pollatsek, & Konold, 1983) in which several participants responded to probabilistic statements as if those statements were true with certainty. In the study reported here, evidence for errors resulting from a nonstandard interpretation of probability was sought by examining participants' verbalizations as they reasoned about various situations involving uncertainty. On the basis of their statements, a model of nonstandard reasoning under uncertainty was formulated. According to this model, referred to as the *outcome approach,* the goal in dealing with uncertainty is to predict the outcome of a single next trial. For example, many participants given an irregularly shaped bone to roll and asked which side was most likely to land upright appeared to interpret the question as a request to predict the outcome of a single trial. These same individuals tended to evaluate their predictions as being correct or incorrect after one trial. Furthermore, outcome-oriented participants often based predictions on a causal analysis of the situation. Numbers assigned as "probabilities" were used occasionally to gauge the strength of these perceived causal factors. More typically, assigned probabilities served as modifiers of the yes–no prediction, with 50% meaning that no sensible prediction could be made.

The outcome approach differs from formal theories of probability and will be contrasted in particular to the *frequentist* and *personalist* interpretations. To the frequentist, a probability is meaningful only with respect to some repeatable event and is defined as the relative frequency of occurrence of an event in an infinite (or very large) number of trials (Reichenbach, 1949; von Mises, 1957). This is viewed as an "objective" theory because the frequentist regards the probability as referring to an empirical, verifiable quantity. A rival subjective theory is the personalist interpretation (de Finetti, 1972; Savage, 1954), which holds that a statement of probability of some event communicates the degree of belief of the speaker (measured by the amount that would define a "fair bet") that the event will occur.

Though theorists quibble over whether some event ought to be assigned a probability, and over the interpretation of the probability, the various schools generally derive identical probabilities for events they all agree are probabilistic. For example, the probability in coin flipping of the outcome *heads* would be determined as .5 on the basis of the classical interpretation, because the ratio of favorable to total number of equally likely alternatives is 1 to 2. For the frequentist it would be .5 if the limit of the relative frequency of heads approaches .5 as the number of trials approaches infinity. According to the personalist interpretation, different people could validly assign different values to the probability of a particular coin based on their beliefs about factors such as the fairness of the coin, the character of the person doing the flipping, the technique of flipping. In formalizing a personalist view, however, theorists have included various adjustment mechanisms requiring the revision (or "calibration") of initial probabilities given new information about the actual occurrence of the event. Savage (1954), for example, advocated the use of Bayes's theorem to revise initial beliefs. Given enough data about the frequency of occurrence of heads with a particular coin, subjective probabilities are thus constrained to converge on the frequentists' limit. It is at this level that the outcome approach is contrasted to formal theories of probability. That is, the outcome-oriented individual does not regard frequency information as relevant in cases where formal theories all would agree that it is.

OVERVIEW OF STUDY

In this study, participants were interviewed on two occasions. In Interview 1, a set of questions dealing with various aspects of probability was given to 16 undergraduates. Videotapes of these interviews were analyzed, and aspects of students' reasoning that were at variance with formal probability theory were identified. Proceeding on the assumption that there were logical connections between various statements that students made, a two-feature model of their reasoning, the outcome approach, was developed. Responses that could be regarded as indicators of reasoning consistent with features of the outcome approach were then coded. On the basis of this coding, a score was generated for each student reflecting the degree of adherence to the outcome approach. Interview 2 was then conducted to test the predictive validity of the outcome approach. Three quarters of the group returned for Interview 2 and were given another set of problems for which specific predictions had been made on the basis of their performance in Interview 1. Together, these data suggest that the account given by Tversky and Kahneman (1983) is incomplete, that some people arrive at probabilistic judgments neither according to formal theory nor through judgment heuristics, but via an alternative interpretation of probability for which the objective is the successful prediction of individual trials.

INTERVIEW 1

Method

Participants. Interview 1 was undertaken to identify aspects of people's reasoning that were nonnormative yet were applied consistently to a variety of problems involving uncertainty. Sixteen undergraduate students at the University of Massachusetts at Amherst were interviewed as they attempted to solve word problems involving uncertain outcomes. Students volunteered to participate in return for extra credit in a psychology course.

Problems. The three problems and follow-up questions used in Interview 1 are presented here in abbreviated form. (The problems in their entirety are included in the Appendix.)

> *Weather problem.* What does it mean when a weather forecaster says that tomorrow there is a 70% chance of rain? Suppose the forecaster said that there was a 70% chance of rain tomorrow and, in fact, it didn't rain. What would you conclude about the statement that there was a 70% chance of rain? Suppose you wanted to find out how good a particular forecaster's predictions were. You observed what happened on 10 days for which a 70% chance of rain had been reported. On 3 of those 10 days, there was no rain. What would you conclude about the accuracy of this forecaster?

> *Misfortune problem.* I know a person to whom all the following things happened on the same day. First, his son "totaled" the family

> car and was seriously injured. Next, he was late for work and nearly got fired. In the afternoon he got food poisoning at a fast-food restaurant. Then in the evening he got word that his father had died. How would you account for all these things happening on the same day?
>
> *Bone problem.* I have here a bone that has six surfaces. I've written the letters *A* through *F*, one on each surface. If you were to roll that, which side do you think would most likely land upright? How likely is it that *x* will land upright? [Student is asked to roll the bone to see what happens.] What do you conclude about your prediction? What do you conclude having rolled the bone once? Would rolling the bone more times help you conclude which side is most likely to land upright?

The problems were selected to vary along several dimensions. A diverse set of problems was used not to see how student responses might vary over problem-type, as in the case of Nisbett et al. (1983), but rather to search for student response-types that persisted across different problems. One indication that the problems do differ is given by their ranking according to the criteria mentioned by Nisbett et al. The bone problem involves (a) a reasonably clear sample space and evident repeatability of trials, (b) easily identified chance factors, and (c) strong cultural prescription toward viewing the phenomena statistically. The misfortune problem exemplifies what Monod (1972) referred to as "absolute coincidence," which involves the convergence of independent chains of events. (A possible dependency between the car accident and the late arrival at work was intended, but all other relations among the events were intended to suggest independence.) The misfortune problem involves (a) both an ambiguous sample space and trial unit, (b) unclear chance factors in that there is no obvious randomizing mechanism, and (c) a situation that is frequently not viewed statistically (Falk, 1981). The weather problem (a) is intermediate in the clarity of sample space and repeatability of trials, (b) involves nonapparent chance factors, and (c) is intermediate on cultural prescription to view statistically. One feature that is consistent across problems is that the elementary outcomes are not equally likely.

The problems also involve a variety of tasks. In the bone problem, students are asked to identify and then generate an estimate of the associated probability of the most likely outcome. In the weather problem, they are asked to interpret a numeric probability, and in the misfortune problem, to provide an explanation of an event. The task of generating probabilities fits into task categories developed by Howell and Burnett (1978). But the tasks of "explaining" an event (categorizing it as a product of chance or of specific causes) and of explicating the meaning of a probability do not appear in their taxonomy.

Procedure. I interviewed students individually in a session lasting approximately 1 hr. Students were instructed that they would be given several problems requiring reasoning about situations involving uncertainty. They were told that the particular answers they gave were of less interest than the reasoning that led to the answer. Accordingly, they were instructed to "think aloud" as they attempted to solve each problem, verbalizing their thoughts as they occurred rather than attempting to reconstruct them at some later time. A felt pen and a pad of paper were provided for their use. Students were informed that the interview would be videotaped, and the recording equipment was in full view.

The problems were presented orally. Two orders of presentation were used, the order being alternated on each successive interview. Order A was the sequence: weather, bone, misfortune. Order B was the reverse sequence.

The majority of probes used during the interview consisted of requests to repeat a statement and reminders to verbalize. However, unplanned probes were used occasionally in an attempt to further elucidate students' thinking.

Results and Discussion

A qualitative analysis of the interview protocols suggested that a subset of students was reasoning according to a nonnormative, yet coherent, belief system. This system can be characterized as involving two general features: (a) the tendency to interpret questions about the possibility of an outcome as requests to predict the outcome of a *single trial* and (b) the reliance on *causal* as opposed to stochastic explanations of outcome occurrence and variability.

To give an initial impression of this belief system, hereafter referred to as the outcome approach, two composite interviews are juxtaposed in Table 1. On the left is a prototype of the outcome approach; on the right, a prototype of a frequency interpretation. These prototypes assemble excerpts from several students (as noted at the beginning of each excerpt) and should be regarded as ideal characterizations. Only a few of the students' individual protocols closely resemble either prototype.

In the remainder of this section, the two features of the outcome approach are described more formally and exemplified with reference to numbered excerpts in Table 1.

Predicting Single Trials

Two types of statements indicated that some students perceived their goal as predicting outcomes of single trials. These statements consisted of (a)

TABLE 1
Comparison of Outcome and Frequentist Responses

Outcome Approach	*Frequency Interpretation*
Weather Problem	
I: What does it mean when a weather forecaster says that tomorrow there is a 70% chance of rain?	
(1) *S5:* What it means is they can see all these cloud patterns forming and moving into a particular area, but they're not as dense as, say, a hurricane where you can absolutely predict where it's going to go. 100%—that means it was a total cloud thing coming over the area.	*S4:* 70% means that the chances that it will rain are 7 out of 10, according to him.
I: What does the number, in this case the 70%, tell you?	
(2) *S6:* Well, it tells me that it's over 50%, and so, that's the first thing I think of. And, well, I think of the half-way mark between 50% and, say 100% to be like, well, 75%. And it's almost that, and I think that's a pretty good chance that there'll be rain.	*S4:* Well, it says that there's a 30% chance that it isn't going to rain.
I: Suppose the forecaster said there was a 70% chance of rain tomorrow and, in fact, it didn't rain the next day. What would you conclude about the statement that there was a 70% chance of rain?	
(3) *S12:* Well, that maybe they just fouled up. Or during the night, the precipitation or something changed in a different direction because of other outside factors.	*S4:* Well, on the basis of just the sample, I think an unrational response would be that the prediction was wrong. But, in fact, 30% is a pretty good probability that it's—it's not miniscule that it's not going to rain.
I: Suppose you wanted to find out how good a particular forecaster's predictions were. You observed what happened on 10 days for which a 70% chance of rain had been reported. On 3 of those 10 days there was no rain. What would you conclude about the accuracy of this forecaster?	
(4) *S3:* Well, I suppose he probably should do better than that. I assume they're trying their best. They're not trying to feed you wrong information.	*S2:* He was exactly right. Seven out of 10 times is 70%. And he concluded 70% chance of rain all 10 times. So—70% of all the time.
I: What should have been predicted on the days it didn't rain?	
(5) *S12:* Well, he could either have said that there's a chance that it might rain rather than being more definite, or just said "mild," you know, "some clouds," or something like that rather than being specific.	

(Continued)

TABLE 1 *(Continued)*

Outcome Approach	*Frequency Interpretation*
Misfortune Problem	
I: I know of a person to whom all of the following things happened on the same day. . . . How would you account for all these things happening on the same day?	
(6) *S5:* I'm trying to figure out if the order you gave me was the order that they happened, or if his father died—or he went out to a family restaurant with his family and they got food poisoning, and because he was sick, while he was driving he smashed up the car. His father died in the accident, and he was on his way to work so he was late.	*S2:* It's arbitary, somewhat. It just occurred. I don't see any other way I could explain how they all occurred on the same day. I could see how if the guy totaled his car, he'd probably be late for work. Even though it's unlikely to occur, like if it only happens 1 in 1,000 times, if you live 1,000 days the odds are it's going to happen to you. So even though it's unlikely for an everyday occurrence, when you consider all the days that you live, it's not so unlikely.
Bone Problem	
I: If you were to roll this, which side do you think would most likely land upright?	
(7) *S9:* Wow. If I were a math major, this would be easy. *B* is nice and flat, so if *D* fell down, *B* would be up. I'd go with *B*.	*S2:* I don't think I could tell you without rolling it. This is not like a die, and I think that there is no way of knowing personally without experimentation. *S4:* I could only give my best guess. I'd have to say *B* up.
I: And about how likely do you think *B* is to land upright?	
(8) *S9:* I wouldn't say it's much more likely. It depends on the roll, I think.	*S4:* I'll give a big bias to *B*. I'll say 33%.
I: So what do you conclude, having rolled it once?	
(9) *S10:* Wrong again. [*B*] didn't come up.	*S15:* I don't conclude anything. Can I roll it again?
I: Would rolling it more times help you conclude which side, if any, was most likely to land upright?	
(10) *S9:* No, I don't know. I think it's difficult to decide which is more likely. I don't see how you really can, just by looking at it. That's my opinion.	*S1:* Oh definitely. I mean that's the only way I could tell for sure. I think the only way with a thing like this is to just keep rolling it and just record the results.

Note. I = interviewer; S1 = Student 1, and so on.

qualitative (yes–no) predictions and (b) right–wrong evaluations of predictions.

Qualitative predictions. In the outcome approach, predictions of single trials take the form of "yes," "no," and occasionally "I don't know" decisions of whether a particular outcome will occur. This contrasts with the frequency interpretation in which the objective typically is to predict a global index of the entire sample, such as the mean or percentage of some outcome in a series of trials. Four students translated the statement "70% chance of rain" into the more definitive qualitative statement, "It's going to rain." This translation was usually accomplished by using the range of 0% to 100% as a decision continuum, with 0% meaning *no,* 100% meaning *yes,* and 50% meaning *I don't know.* Intermediate values were ultimately associated with one of these three anchor or decision points according to a vague and variable proximity criterion. Thus, 70% was considered sufficiently above 50% to warrant identification with 100%, or *yes,* with perhaps some associated expectation of error (see Excerpt 2). Given this qualitative (yes–no) interpretation of the probability range, 50% was viewed by 3 students not as a predictive forecast, but as an admission by the forecaster of total ignorance about the outcome. For example, Student 9 replied:

> It's not 100% chance and its not 50–50, so he's not guessing. If he said 50–50 chance I'd kind of think that was strange . . . that he didn't really know what he was talking about, because only 50–50—"it might rain or it might be sunny, I really don't know."

Evaluation of predictions. In both the bone and weather problems, several students indicated that a probability value was either right or wrong after the occurrence of a single trial. This evaluation suggests that they perceive the goal in such situations as correct prediction of single trials.

In the weather problem, a situation was posed in which no rain fell on a day for which a 70% chance of rain had been estimated. Asked what they would conclude about the accuracy of the statement that there was a 70% chance of rain, 6 students responded that the statement must have been incorrect (see Excerpt 3). Students were also questioned about the accuracy of a forecaster who had predicted 70% chance of rain for 10 days, when in fact no rain was recorded on 3 of the 10 days. Theoretically, 7 days of rain out of 10 is the most likely outcome given an accurate 70% forecast on each day. Three of the students' responses were consistent with this reasoning. Nine students concluded, however, that the forecaster was only "pretty accurate," suggesting that there was room for improvement (see Excerpt 4). Four students expressed a conflict over whether the forecaster was perfectly accurate or not. At the heart of this conflict was the question of whether the forecaster is trying to formulate (a) an accurate prediction of the relative frequency of rainy days or (b) a decision about whether or not it will, in fact, rain. Student 8 concluded:

> Well, he's looking at an individual day—particular day—and he's setting up percentages on one day. And you can't really extend that to an amount of time, I don't think.

In the bone problem, students were first asked to make an initial guess regarding which side of the bone was most likely to land upright. After stating a probability that the chosen side would land upright, they were asked to roll the bone. Nine students remarked that their guess was either right or wrong having observed the result of one trial (see Excerpt 9).

Students' statements from both the bone and weather problems suggest that a subset of students encoded requests for probabilities as requests for a decision of which alternative would occur on a particular trial. Once the trial had been conducted, these predictions were retrospectively evaluated as having been right or wrong. When probabilities were provided, as in the weather problem, they were not interpreted as probabilities per se, but as values that could be used to formulate a yes-no decision.

Predicting Outcomes From Causes

For each problem, several students made statements indicating that they generated or interpreted probability estimates via a causal analysis of the problem situation.

Weather problem. For the weather problem, students were asked to explain the meaning of the number in the proposition, "There is a 70% chance of rain." Four of them suggested that the 70% was a measure of the strength of a factor that would produce rain (e.g., 70% humidity or 70% cloud cover; see Excerpt 1). Three students used causal explanations to account for the nonoccurrence of rain given the forecast of 70% chance of rain (see Excerpt 3).

Misfortune problem. Eight students gave other-than-chance explanations of the several low-probability events in the misfortune problem. Six students tried to embed all the events in a causal sequence so that each could be seen as resulting directly from a preceding event (see Excerpt 6). Five students relied on explanations that involved causal agents such as God or the stars.

Bone problem. In the bone problem, 5 students expressed reservations about whether additional trials would be helpful in determining which side was most likely to land upright (see Excerpt 10). Three of these individuals suggested that more reliable information could be obtained from careful inspection of the bone than from conducting trials. Three students did not use the provided results of 1,000 trials to predict the results of 10 trials. Eight students attributed variations among trials to the way the bone was rolled.

It needs to be stressed that a formal probabilistic approach does not necessitate the denial of underlying causal mechanisms in the case of chance events. Hypothetically, one can imagine describing the last in a series of 100 tosses of a fair coin in sufficient detail that it could be seen to be determined by preceding events. In practice, however, a causal description is often seen as impractical if not impossible (e.g., von Mises, 1957, pp. 208-209). Accepting a current state of limited knowledge, a probabilistic approach adopts a "black-box" model according to which underlying causal mechanisms, if not denied, are ignored. The mechanistic model is not abandoned in the outcome approach. The goal of predicting the results of individual trials in a yes-no fashion seems to imply the possibility of determining beforehand the results of each individual trial.

Degree of Adherence to the Outcome Approach

A variety of responses to the problems used in Interview 1 suggest that some individuals occasionally employ a nonstandard approach to probability. The salient features of this approach are (a) predicting outcomes of single trials, (b) interpreting probabilities as predictions and thus evaluating probabilities as either right or wrong after a single occurrence, and (c) basing probability estimates on causal features rather than on distributional information.

Table 2 summarizes the statements made by each student that were indicative of the features just listed. Brief descriptions of the statement types are listed down the left of the table, grouped by feature and, within feature, by problem. The checkmarks (✓) encountered reading down the table under a student's number indicate the outcome-oriented statements made by that particular student.

Inspection of Table 2 indicates that the outcome approach is not a belief system that individuals either do or do not hold. All but 2 students gave at least one outcome-oriented response. Some responses were made hy a majority of students (e.g., right-wrong evaluation in the bone problem), whereas others were made by only a fifth of them (e.g., "50% means anything can happen"). Rather than viewing the outcome approach as a discrete category, it is viewed here as a set of beliefs that individuals hold to differing degrees.

The outcome scores at the bottom of Table 2 were determined by adding

TABLE 2
Outcome-Oriented Responses: Interview 1

		Student Number[a]															
Problem	*Response Category*	*1**	*2**	*3*	*4**	*5*	*6*	*7*	*8**	*9*	*10*	*11*	*12*	*13*	*14**	*15*	*16**
	Single-Trial Feature: Evaluative Response																
Bone	Prediction, right/wrong	✓		✓		✓				✓	✓	✓	✓	✓			✓
Weather	Forecaster right/wrong			✓						✓		✓	✓	✓		✓	
	7/10 → pretty accurate			✓	C[b]	C	✓		C	✓	✓	C	✓	✓	✓	✓	✓
	Single-Trial Feature: Qualitative Interpretation																
Weather	50% < 70% < 100%	✓		✓		✓	✓		✓	✓			✓	✓	✓	✓	✓
	70% → rain					✓							✓	✓		✓	
	50% → anything can happen								✓	✓					✓		
	Causal Feature																
Bone	Additional trials no help						✓			✓	✓	✓	✓				
	Ignore data from 1,000 rolls			✓									✓	✓			
	Predict via physical features						✓			✓			✓				
	Variability due to "the roll"			✓		✓			✓	✓		✓	✓	✓		✓	
Weather	70% → strength of causes			✓		✓				✓			✓				
	No rain → change of weather								✓	✓			✓				
Misfortune	No mention of chance			✓								✓		✓			
	External, controlling force			✓				✓				✓	✓	✓			
	Internal, causal connection					✓		✓		✓		✓	✓		✓		
Total outcome score		2	0	9	0	6	4	2	4	11	3	7	13	9	4	5	3

[a] Students with starred numbers had had a college or high-school statistics course.
[b] C indicates conflict between "good" and "perfect" accuracy.

the number of outcome-oriented statements checked for each student. These scores serve in Interview 2 as a measure of an individual's degree of adherence to the outcome approach. Possible scores ranged from 0 to 15, with higher scores indicating an outcome orientation. The median for the 16 students was 4.17, with actual scores ranging from 0 to 13. The students with low outcome scores in general gave responses that were consistent with a formal interpretation of probability. These responses were not coded, but are exemplified in Table 1. Students who had taken a statistics course (as noted in Table 2) tended to receive lower outcome scores than those who had not, with means of 2.17 and 6.9, respectively.

INTERVIEW 2

Data from Interview 1 were analyzed qualitatively to develop a model, the outcome approach, of nonstandard reasoning under uncertainty. A second set of interviews with the same participants was conducted to (a) establish that, within individuals, outcome-oriented responses tend to be consistent across problems and (b) test the power of the model to predict specific responses that had not been observed in Interview 1.

Method

Participants. Twelve of the original 16 students returned to participate in the follow-up interviews. Outcome scores from Interview 1 for these students ranged from 0 to 13, with a mean of 5.3 and a median of 4.5. The other 4 students, who could not be located, had mean and median outcome scores of 4.5 and 3.5, respectively. Approximately 5 months had elapsed between Interviews 1 and 2. The length of this interval resulted from the time required to analyze the data from Interview 1 and develop the problems and predictions for Interview 2. During this time Students 1, 2, and 6 had enrolled in an introductory statistics course and were near completion when interviewed. Students 1 and 2 had taken statistics previously in high school.

Problems and procedure. Four problems were employed. The cab problem has been used in previous research (Kahneman & Tversky, 1972). The three remaining problems were developed and then standardized in 14 pilot interviews. All four problems are presented here in abbreviated form in their order of occurrence in the interview. The problems are presented in their entirety in the Appendix.

> *Cab problem.* [Student is asked to read the cab problem aloud.] "A cab was involved in a hit-and-run accident at night. Two cab

companies, the Green and the Blue, operate in the city. You are given the following data:

1. 85% of the cabs in the city are Green and 15% are Blue.
2. A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs, half of which were Blue and half of which were Green, the witness made correct identifications in 80% of the cases and erred in 20% of the cases.

What is the probability that the cab involved in the accident was Blue rather than Green?"

Bone 2 problem. Last time you were asked which side of this bone you thought would most likely land upright. Do you remember which side you concluded? [The bone is held far enough away so that the labels cannot be read.] I'm going to ask you the same question again. And to give you something to base your answer on, I'll offer you any one of the following pieces of information. [Student is shown the list as the interviewer reads the items.]

1. A measure of surface area of each side.
2. The results of 100 rolls made by 16 people.
3. The results I got in 1,000 rolls.
4. A drawing of the bone showing the center of gravity.
5. The bone to look at.
6. The results of your last 10 rolls.

Painted-die problem. I have here a six-sided die. Suppose I painted five of the surfaces black and the other one white. If I rolled the painted die six times, would I be more likely to get six blacks or five blacks and one white? If I rolled it 60 times, how many times would you expect the white surface to come up?

Modeling problem. Would there be a . . . way that we could make a model of the bone so that, instead of rolling the bone, we could pick something out of a container and get the same kind of results? [If a student cannot generate a model, four possible models are suggested in succession, and the student is asked to comment on their appropriateness. When, and if, the student settles on a model of the bone, he or she is asked the following questions:] Suppose I rolled the bone 100 times and kept track of what I got, then I drew 100 times from this can filled with the labeled stones. If I showed you the results from

> both, could you tell from looking at the results which I got from rolling the bone and which from drawing from the container? In those 100 trials with the bone and the container, do you think with one of those I'd be more likely than with the other to get no *E*s? Do you think I'd be more likely with one of those to get more *D*s in 100 trials than with the other?

Initial instructions to students were similar to those given in Interview 1. They were told that they would be given several problems involving uncertain outcomes. They were reminded to "think aloud" and to use the pen and paper for any figuring they might want to do. All the problems except the cab problem were presented orally. The entire interview required approximately 40 min.

Results and Discussion

The four problems used in Interview 2 were designed to determine whether the responses of outcome-oriented individuals to another set of questions could be predicted. To test these predictions, scores based on performance in Interview 2 were correlated with the outcome score that summarized students' performance in Interview 1.

Although the full rationale for choosing these four problems is made clear in the subsequent discussion, I summarize some of their features here. In prior research using the cab problem, participants had made statements consistent with the single-trial prediction feature of the outcome approach. The cab problem was selected for Interview 2 as an independent measure of the consistency of this feature over problems and sessions. In the bone 2 problem, students were again asked to predict outcomes of rolling the same bone used in Interview 1. A different set of probes was used to determine whether estimates were being generated primarily from frequency information or from physical features of the bone. Given that the unit of analysis in the outcome approach is the single trial, it was predicted that outcome-oriented students would solve the painted-die problem by first imagining the results of each individual trial and then concatenating these results to obtain the solution for six trials. Because black is the best guess for each of the six individual trials, it was predicted that outcome-oriented students would believe that six blacks are more likely than the normative solution of five blacks, one white. Finally, the modeling problem was designed to test the validity of the causal feature. It was predicted that outcome-oriented students would not believe that an urn model could be constructed to simulate the results of rolling the bone, because salient causal features would be altered.

In the remainder of this section, each problem and the associated predictions are discussed in turn. After specifying the predictions made prior to conducting the interviews, correlations between performance in Interviews 1 and 2 are reported, and then selected excerpts from the interviews that pertain to the predictions are discussed.

Cab problem. The cab problem (originally used by Kahneman & Tversky, 1972) has been used to study individuals' reluctance to take into account base rates (in this case the relative number of the two colors of cabs) in the formulation of probability estimates. Well et al. (1983), using an interview format, reported that many participants believed they were being asked not the probability that the errant cab was blue, but whether it *was* blue. In addition, numeric answers that participants were asked to provide in many cases seemed to be only loosely based on the numbers given in the problem. These observations are similar to students' responses to the bone and weather problems.

Given that the outcome approach describes a general orientation to uncertainty, those who responded in an outcome-oriented fashion in Interview 1 should respond in a similar way to the cab problem. Specifically, it was predicted that outcome-oriented students, as defined by higher outcome scores in Interview 1, would be more likely to:

1. Ask whether a number was required in answering the question of the probability that it was a blue cab.
2. Encode the question, "What is the probability . . . ?" as the question, "What color was the cab?" (this encoding being indicated by qualitative responses such as, "I think the cab was blue").
3. Base a numeric answer on a "loose" or qualitative interpretation of the evidence they thought relevant.

Coders for this problem (and for the other three problems) were myself and a graduate student who was blind both to the nature of Interview 1 and to the hypotheses being tested. Interrater reliability for coding the three categories of the cab problem was estimated by correlating the set of ratings of the two coders ($r = .759$). The scoring rule applied was that both coders had to agree that a particular category of statement had been made for it to be counted.

The three response categories for the cab problem are listed down the left of Table 3. Students are ordered across the top of the table according to their outcome scores on Interview 1. The Interview 1 outcome scores are provided below the student numbers, with scores increasing toward the right of the table. The checkmarks indicate the students who made each particular response type. For example, Student 1 (S1), who had an outcome score of 2 on Interview 1, made both a "number inquiry" and a "qualitative

TABLE 3
Outcome-Oriented Responses: Interview 2

	Student Number											
	2	*1*	*7*	*16*	*6*	*8*	*15*	*5*	*11*	*3*	*13*	*12*
	Outcome Score: Interview 1											
Response Category (by Problem)	*0*	*2*	*2*	*3*	*4*	*4*	*5*	*6*	*7*	*9*	*9*	*13*
Cab problem												
Number inquiry		✓					✓	✓	✓		✓	
Qualitative statement		✓					✓					✓
"Loose" numeric answer										✓	✓	✓
Bone 2 problem												
First choice not frequency												✓
Second choice not frequency			✓				✓	✓		✓		✓
Predicted from physical properties								✓		✓		✓
Statisticians use physical properties							✓	✓	✓	✓	✓	✓
Painted-die problem												
6 blacks in 6 trials			✓	✓			✓		✓	✓	✓	✓
< 10 whites per 60 trials			✓	✓			✓			✓	✓	✓
$p(B) > 5/6$							✓			✓	✓	✓
Modeling problem												
Reject urn model of die										✓	✓	
Model of bone not generated after probe			✓		✓		✓			✓	✓	✓
Reject 100-labeled-stones model of bone			✓		✓		✓			✓	✓	✓
Reject trial-results model of bone					✓					✓	✓	
Outcome score: Interview 2	0	2	5	2	3	0	9	4	3	11	10	11

statement" in response to the cab problem. To the extent that outcome scores predict responses in Interview 2, the checkmarks in Table 3 should become more numerous toward the right of the table.

To quantify the relationship between outcome scores and performance on the problems used in Interview 2, student scores were generated on each Interview 2 problem by adding the number of checked response categories. These problem scores were then correlated with the outcome scores from Interview 1. The correlation between scores on the cab problem (which could range from 0 to 3) with the outcome scores from Interview 1 was $r = .586$ ($p < .025$, one-tailed).

An examination of various students' statements suggests that the percentages given in the cab problem were not interpreted as probabilities but rather were used to formulate a definitive statement about the color of the hit-and-run cab. Given that the goal of the outcome approach is to determine what will or did occur, the question concerning probability is translated into the question "What happened?" as indicated in the following response (from Student 1):

> So you want to know if I think that's right—if it was blue. Well, I would say it would be blue rather than green—just the fact that this really isn't important—the 85% are green, 15% are blue. I mean there are still a substantial amount of blue cabs out there. But the fact that the guy said—well, the court said that "in 80% of the cases you identified the right color." And the guy said he saw blue. He doesn't say "I think I saw blue." He says, "I saw blue." So I would go with blue.

In the cab problem, students were asked specifically for the probability that the cab was blue. A query of whether a number was required was considered consistent with the yes–no feature of the outcome approach. When students asked if a number was wanted, I hesitated to allow them to clarify the question, and then if they did not continue, I asked what the alternative was to giving a numeric answer:

> *S11:* Let's see. Am I looking for a number as opposed to like—Am I looking to say, "It's 80% probability that it was a blue rather than green?" Is that what I'm—
>
> *I:* What's the other option? How else would you prefer to give that?
>
> *S11:* "Sure, it could have been a blue cab." [Laughter] No,—just that it would have been a strong—it was more likely as opposed to less

> likely. Kind of like this fit in. More positive as opposed to a definite number positive.

Central to the goal of specifying what will happen or did happen is the focus on single trials: Questions of uncertainty are viewed as pertaining to a particular event as opposed to a set of events. Student 5 justified ignoring the base-rate information on the grounds that the occurrence of a particular event was at issue, and that information regarding a class of events was irrelevant:

> It really doesn't matter how many cabs there are in the city. What you're thinking about is this one particular cab, whether it was blue or green. And since the guy was usually right, he's probably right.

As suggested in this excerpt, the witness identification can be seen as applying to the individual event (the color of the errant cab) in a way that the base-rate information cannot. Using the base rates would seem to require regarding the particular accident as one of a set of accidents involving the two cab companies. To the outcome-oriented individual, this is not relevant to the question; what matters is this particular accident. It is evident in the preceding and following excerpts that the witness identification is not viewed as one of a class of similar identifications. Rather, the outcome-oriented individual may assign the attribute "pretty reliable" to the witness and thus to the witness's identification of the errant cab's color based on the accuracy data collected by the court. It may be in the process of assigning this attribute that students "let go" of the specific meaning of the 80% and give a confidence value for their belief that the cab was blue that is only loosely based on the 80% estimate of the witness's accuracy.

> *S8:* And since his visibility was pretty clear, and just on that—I'm not even taking these numbers so much as just, you know, conceptualizing it. Since he saw it was blue, and there's more of a chance that he's right as seeing it as blue, that he saw it correctly. So I'll say that.

> *S3:* 80% just because he had—his percentage correct before was 80%, so it makes sense that he, probably—chance 80% that he got it right this time.

> *I:* Okay.

> *S3:* Maybe better.

I: Can you explain why you think it might be better than that?

S3: Well, because more than not he got them right when they tested him before. So that's why it would be possible that he'd be more than 80%.

S13: Yeah—that he did guess, more than he didn't, the right colors. So I'd go with the blue. I'd say that it was a blue one.

I: And how about just an estimate of what the probability would be, or a guess.

S13: I want to say just 80 . . .

I: Is that 80 based on this [pointing to 80% witness accuracy]?

S13: No. I'm just trying to find—I'm just trying to think of something that's closer to 100—like over to more of a chance that it happened.

Bone 2 problem. Given another opportunity to decide which side of the bone was most likely to land upright, it was predicted that outcome-oriented students would prefer to consider the physical features of the bone rather than frequency data. It was also predicted that, when asked how a statistician would determine the probabilities associated with each side of the bone, outcome-oriented students would express the belief that various physical features of the bone would be taken into account.

Scores for this problem could range from 0 to 4 based on the following four response categories:

1. Frequency data was not the first-choice information for determining the most likely result of rolling the bone.
2. The second choice was not frequency data.
3. Physical properties of the bone were used in predicting the results of 10 trials.
4. It was believed that a statistician would consider physical properties in determining probabilities associated with rolling the bone.

Performance on this problem is summarized in Table 3. The interrater reliability for coding students' statements with respect to these categories was 100%. The correlation between scores on the bone 2 problem and the outcome scores from Interview 1 was $r = .782$ ($p < .005$, one-tailed).

As predicted, outcome-oriented students were more likely to believe that a decision about the probabilities of various sides of the bone landing upright should be arrived at by considering the physical features of the bone. One hypothesis to account for why they preferred a physical to a statistical analysis is that the physical features of the bone might be viewed as a more stable source of evidence when compared with frequency data, which can fluctuate from sample to sample. This seemed to be the rationale given by Student 12 for basing predictions on an inspection of the bone. Asked why she thought the data from 100 rolls were unimportant, she replied,

> *S12:* Well, because what they did may not be—it's sort of chance, you know, that happened. If the same 16 people did the same 100 rolls, it would probably be different the second time. It just doesn't seem a very specific kind of statistic.
>
> *I:* And why do you think it would be different?
>
> *S12:* Things change. I don't think anything duplicates itself exactly the second time.
>
> *I:* How about the results I got in 1,000 rolls?
>
> *S12:* Yeah, that too is kind of iffy. If you did the same thing over again, plus a second 1,000 rolls—I mean, you could go on for 2,000 rolls or whatever, and I don't know if it really would tell you much. Then again, I could be wrong.

A second hypothesis to explain why a physical analysis might be preferred is that physical properties may be viewed as causal agents of what one wants to predict, whereas frequency information is not. The interviews, however, provided no compelling evidence that this was the case. One student did express the belief in Interview 2 that the physical properties were "real evidence" in contrast to frequency data. Asked to explain how she decided that *D* was the most likely side, she responded:

> *S3:* Well, just 'cause it's flatter on the underside, so it's more likely to land on that side than it would on any other place.
>
> *I:* Are you using this information at all [the results on her last 10 rolls]?

> *S3:* Maybe a little, yeah. I suppose. Well, I looked first and thought that was reasonable, so . . .

Asked how a statistician would determine the probability, she first mentioned surface area. Asked if they would use anything else, she replied:

> *S3:* Well, they would probably make rolls themselves and see how it comes up. But I don't know if they would use that for real evidence or whatever.
>
> *I:* You feel like the results of what you got isn't real evidence?
>
> *S3:* Well, yeah. It has some. But there must be some, you know, like measuring the sides, and that must be a little more precise than my rolls.

This last statement suggests that she regards the properties of the bone as more valid evidence, because it is easier for her to think of them as being measured precisely.

Students 5, 11, and 13, who also considered features of the bone to be important for determining probabilities, suggested that they should be used in conjunction with, rather than to the exclusion of, frequency data:

> I'd take number 3 [results of 1,000 rolls], and I'd look at each surface of the bone that had come up and compare it to the number of times it had gotten up and see why it had so I could decide whether or not the results were accurate, according to the shape of the bone. (Student 5)

In predicting 10 rolls, Student 5 inspected the bone carefully to decide how he would allot predicted frequencies to *B* and *C* because, according to him, they were so close in frequency of past occurrence. His explanation of how a statistician would estimate probabilities was consistent with the approach he had employed:

> A statistician would count a great deal of weight to the center of gravity and how it related and, taking your results [from 1,000 rolls], would come up with a bunch of statistics that would probably reflect fairly accutately your results, with perhaps some modification according to what he thought the structure of the bone gave out.

Student 11 used only frequency data to make predictions about the bone, but expressed the belief that a statistician would, in a "joint effort," supplement these with an analysis of physical properties:

> *S11:* 'Cause you'd roll the bone and get a rough idea of the probabilities, whatever they are—yeah, probabilities—and take it to have it analyzed to figure out if, structurally, you can understand why these—you know. You assign these particular values to each face, and then through comparing both, just—
>
> *I:* But I might want to modify what I had got rolling it?
>
> *S11:* Yeah. It's just kind of like added significance, or not significance—added sureness, or whatever—belief in your percentages.

In summary, the tendency to view physical properties of the bone as important in the determination of probabilities of the various landing orientations is correlated with measures of the outcome approach obtained from Interview 1. Physical properties appear to be regarded as information at least on a par with frequency data in making predictions.

The correlations between performance on Interview 1 and the first two problems of Interview 2 suggest that students' outcome-oriented responses were consistent over time and problems. The last two problems were designed to elicit responses that had not been observed in Interview 1 but that would be consistent with the outcome approach. Thus, predicted performance on these two problems provides more compelling evidence of the validity of the outcome approach.

Painted-die problem. In the painted-die problem, students were first presented a die and then six stones, both of which consisted of five elementary outcomes of one type (black) and one of another (white). They were asked to predict whether in six trials they would be more likely to observe five blacks and one white, or six blacks. Theoretically, the former is more likely, the probability of exactly five blacks being .402, the probability of six blacks being .335.

It was not expected that any student would be able to generate the binomial expansion to compute these probabilities. It was expected, however, that knowing that the probability of white being rolled was 1 in 6 might allow them to infer that on average they could expect to get one white in six trials, which is also the modal outcome. Even failing this line of reasoning, one would predict on the basis of the representativeness heuristic that people would believe five blacks to be the more likely outcome because it looks more like and, in this case, is identical to the population distribution. Kahneman and Tversky (1973) reported results on a similar problem involving drawing cards with replacement from a deck in which 5/6 of the cards were marked *X* and the remaining 1/6 were marked *O*. They found that 87% of their participants judged five *X*s and one *O* to be more likely than six *X*s.

In contrast, it was predicted that outcome-oriented participants would regard six blacks as the more likely outcome. In the outcome approach, the primary unit of analysis is the individual trial. Application of the representativeness heuristic in this problem requires a focus on predicting the *sample* result rather than the individual trial results. Given a probability value, the outcome-oriented individual arrives at a prediction of an *individual* trial by deciding which yes–no or I-don't-know decision point is closest to the probability value. Thus, rather than viewing 5/6 as a value related to the expected relative frequency of blacks in randomly drawn samples, it was predicted that outcome-oriented individuals would interpret 5/6 qualitatively, giving it the approximate meaning, "The next trial will almost certainly result in a black." When asked to predict the most likely outcome for six trials, rather than using 5/6 to form an expectation for the sample of six trials, they may arrive at a prediction by concatenating their expectations of each of the six trials. Because this prediction is more qualitative than quantitative, it was expected that outcome-oriented students would more frequently say that six blacks are more likely, and that they would also believe that the ratio of blacks to whites over a larger series of trials will remain above the normative value of 5:1.

Scores for the painted-die problem had a possible range of 0 to 3 based on the following three categories:

1. Six black stones were judged as more likely than five blacks and one white.
2. Fewer than 10 white stones were expected in 60 trials, or, on average, more than 6 trials were required to get a white.
3. The probability of a black on the first trial was estimated to be above 5/6 or above 84%.

Individual performance is noted in Table 3. Interrater reliability for coding the painted-die problem was 100%. The correlation between scores on this problem and outcome scores from Interview 1 was $r = .616$ ($p < .025$, one-tailed).

Excerpts from the interviews indicated that, as suggested, students solved the problem by imagining a single trial for which the probability of black is overwhelming, and then extended this prediction over trials to arrive at the conclusion that six blacks was the more likely outcome.

Student 7 initially stated that the probability was 5/6 for a black. Later he stated that 6 blacks was more likely than 5 blacks, 1 white, and that 10 or fewer whites would occur in 60 trials:

> Well, I think it's—the white's there, but—I'm not exactly sure what I'm trying to say. Just because the odds are always the same. There's only one of them in there. So even though it's six rolls and there's six things in there, there's only one or the other that's going to come up each time. And that—chances are better than five to one, one of the five blacks is going to come up.

Similar reasoning is demonstrated by Student 15.

> Because it's a higher probability of getting a black side because there are more black sides, and so there's more probability that when you roll it, you're going to get a black side instead of that one white side.

Student 3 combined the "more blacks" rationale with the reasoning that the sampling-with-replacement procedure does not guarantee white:

> Probably more likely to get all black just 'cause—I don't know what percentage, but most of the die is black, so it's going to come up on that side. 'Cause you're not going to roll it on a different side each time you roll it, so that it's bound to come up one of those six rolls. So it probably would be black on all of them.

Student 5 believed that rolling six dice at once would result in five blacks, but that rolling the same die six times would result in six blacks:

> Well, each roll is a separate entity. You roll it, and a side will come out. You don't roll all six at one time. So likelihood is that each time it comes out, the side that has the dominant color, which is black, is the color that'll come out.

He finally rejected this reasoning, favoring five blacks in both cases. His initial response, however, provides a good example of what is being regarded as the outcome approach to this problem—that of imagining the results of one trial as almost certainly being black, and, by extending this qualitative judgment, concluding that six blacks are more likely over six trials. It is especially significant that this student began thinking differently about the problem when he imagined all six trials occurring at once, changing his focus from six single trials to a set of trials. (A similar belief in a difference between flipping one coin repeatedly and several at once was defended by the 18th-century mathematician, D'Alembert. For an interesting account of this and other of D'Alembert's unconventional beliefs about probability, see Todhunter, 1949.)

Modeling problem. The modeling problem was designed to test an implication of the causal feature of the outcome approach. According to the outcome approach, frequency data are not considered as reliable for predicting outcomes as are phenomena that are causally related to the outcome. This being the case, it was predicted that outcome-oriented individuals would hold that if the causal features of a setup were altered, outcome frequencies for that setup would change accordingly. In the modeling problem, students were asked if it would be possible to construct an urn model of the bone to generate results that would be indistinguishable from results obtained from rolling the bone. Students had been introduced to the modeling concept in the painted-die problem, where it was suggested that randomly sampling with replacement from an urn containing six identically shaped stones would be the same as rolling a fair die. I assumed that most students would accept this comparison, because the most obvious physical feature—the symmetry of the six sides—was maintained. With an urn model of the bone, however, the important physical aspects of the bone—its irregularly shaped sides and unequal weight distribution—are transformed into unequal numbers of objects that are identical in weight and shape. I predicted that outcome-oriented students, focusing on this difference, would expect the data obtained from conducting trials on the two setups to be distinguishable in some way.

Scores for the modeling problem could range from 0 to 4, according to individual performance with respect to the following four categories:

1. The urn model was rejected in the case of the die.
2. An urn model was not generated in response to the probe, "Is there some container that I could fill with some number of lettered stones that would give results similar to rolling the bone?"
3. The suggestion was rejected that the bone could be modeled by filling a can with 100 labeled stones corresponding in number to a statistician's probability estimates for each side.
4. A can filled with labeled stones corresponding in number to the results of any large number of trials with the bone was rejected as a model.

Interrater reliability for coding with respect to these four categories was $r = .93$. The correlation between these scores and the outcome scores from Interview 1 was $r = .508$ ($p < .05$, one-tailed).

The reasons given by students for rejecting the urn models are congruent with the hypothesis that, in their analysis, important causal features could not be duplicated in the urn models. Students 3 and 13 stated that the urn model was inappropriate in the painted-die problem. They expressed concern not over the corresponding features of the die and stone-filled urn per se, but over the differing sampling procedures in the two cases:

> *S3:* I think maybe the white side of the die would come up more, just 'cause you don't have any control over that [makes an imaginary roll of the die]—well, not that you do with the pieces. . . . You're putting your hand in there and taking out. I just, I don't know why, but I don't think you'd pick the white one as often as the white side of the die.
>
> *S13:* I just think grabbing something out—if you're grabbing it out, I think it would be more probable of being white. I don't know exactly why I'm thinking that way, but with this [die] I just [rolls die]—I don't know, tossing something just seems less of a chance, but picking something out seems more of a chance. You'd think it would be the other way around, though. But I don't know. . . .

In the following excerpts, students explain why an urn model is inappropriate in the case of the bone. That the bone has six sides, uneven surfaces, and is rolled rather than drawn from are all facts mentioned as important differences between it and an urn filled with labeled objects.

> *S3:* Probably be more likely to get no *E*s with the container full of 100 pieces. Just—well, there is a slighter chance that it would come up, and there's six sides. So that's why I think it's more likely to come up on the bone.
>
> *I:* Because?
>
> *S3:* Because there's only six sides . . .
>
> *S6:* Probably it would be more likely to get no *E*s from the bone, 'cause the bone has to stand like that, and it would be easier just sitting in there. They don't have to—it's not like there's anything to do with the way it can stand and stuff like that.
>
> [*D*] might be more likely from the bone. I don't really think you can say, but it just might be just because the *D*s are all mixed up in the can, whereas in the bone, that's the easiest side for it to land on. That's the most—that's the way it stands easiest, so you might get it more times in a row in the bone.

S7: You could easily pick up 100 of them without hitting an *E*. You'd have more trouble tossing the bone so you didn't come up with an *E*.

I: And why is that again?

S7: It just seems like because you're picking them out you could just miss one of the *E*s.

S15: These stones and the die are uniform, and each side is the same—it's the same surface. And this [bone] is all different. So this will affect—the shape of the side will affect the way it's going to roll. Like it would be harder for it to stand up on *E* like that. So you'd have to replicate the little indents and stuff like—so you couldn't make a—you couldn't turn it into six stones or something like that.

The persistence demonstrated by students in insisting that the bone could not be modeled was particularly impressive. The interview probes were designed to give them several opportunities to accept a model. They were given one alternative after another. The independent coder, not knowing the intention in this probing, discreetly noted in two instances that the students had been strongly led to accept a model. The other students were as strongly "led" but insisted repeatedly that the model suggested would not be comparable to rolling the bone. Attending to the physical features as opposed to the resultant frequency data of a chance setup appears to be a deeply ingrained orientation.

Consistency of outcome-oriented responses. As a general indication of the consistency of outcome-oriented responses across the two interviews, the outcome scores from Interview 1 were correlated with overall outcome scores from Interview 2. The latter scores were generated by adding the four problem scores for each subject. The range of possible outcome scores for Interview 2 was 0 to 14. Actual scores, which are provided at the bottom of Table 3, ranged from 0 to 11, with a mean of 5.0 and a median of 3.5. The correlation between outcome scores from the two interview sessions was $r = .797$ ($p < .005$, one-tailed).

As reported in the previous sections, outcome-oriented responses on each of the Interview 2 problems correlated positively with the outcome scores from Interview 1. These correlations, along with the overall correlation between outcome scores from the two interview sessions, suggest that students' nonnormative responses were fairly consistent, both across different problems and a 5-month time interval. In the case of the painted-die and modeling problems, the correlations between Interview 1 performance and problem performance are also indicative of the power of the outcome approach as a predictive model.

GENERAL DISCUSSION

As mentioned in the introduction, it has been suggested that two types of cognitions are available to adults in reasoning about uncertainty. These are (a) formal knowledge of probability theory and (b) natural assessments that become organized as judgment heuristics. Nisbett et al. (1983) suggested that most adults use formal, probabilistic knowledge when reasoning about situations that are clearly probabilistic and have a simple sample space. For situations that are less obviously probabilistic or for which the sample space is less tractable, they fall back on the use of judgment heuristics. This account suggests that, in generating a probability value, the major difference between a formal, probabilistic approach and a heuristic approach is in the *method* of generation. The meaning of that value is presumed to remain normative even though the heuristic method used is not.

My results suggest that this account is incomplete, that some people's responses can best be understood as resulting from an alternative interpretation of probability. This interpretation, the outcome approach, is based on causal reasoning with the goal of predicting outcomes of single trials. It is important to stress that the outcome approach is not meant as an alternative explanation of the particular errors that have been explained by Kahneman and Tversky via judgment heuristics, but as an explanation of a different set of errors. It is also important to mention that, because of the small sample size and homogeneity in age and education of the particular students interviewed, no conclusions can be drawn on the basis of this study about the general pervasiveness of the outcome orientation.

Problem Variables

In this study the outcome approach has been portrayed as a belief that various people hold to differing degrees. As with the use of heuristics, however, features of a problem no doubt can induce the use of the outcome approach among individuals who would apply formal probability in simpler situations. Let us consider two such features.

Most of the problems used in this study involved elementary events that were not equally likely, and this feature might prompt the outcome orientation. Students encountering phenomena like the bone roll (e.g., loaded coins and die) may no longer think of sampling as resulting in a random, chance event and may, therefore, regard a causal analysis as appropriate. In the extreme it could be argued that some students believed, for example, that the bone would always land in the same orientation. Such a belief might be motivated by the observation that, because the bone is not symmetrical, it is no longer "just chance" that is involved. This observation might mislead them to believe not that the distribution of outcome probabilities is nonuniform but that only one outcome is possible. This account suggests that the outcome approach is not an alternative interpretation of probability, but simply a result of believing that loaded coins always land on one side. If individuals held this belief, they would be justified in taking a single-trial approach. Attaching a probability of less than 100% to their prediction in this case could be a valid description of their uncertainty of which side would *always* come up. This account, however, is not supported by the data. In the painted-die problem, students believed that the minority outcome would sometimes occur but still seemed to adopt a single-trial analysis. Also, after rolling the bone two or three times, it was clear that more than one outcome was possible, and yet some students persisted not only in focusing on single-trial predictions, but also in thinking frequency data were not particularly useful as information.

For some individuals the outcome approach may be maintained even in situations involving uniform outcome probabilities such as flipping fair coins. Some preliminary data (Konold, 1988) using items sensitive to both heuristic errors and outcome-oriented responses suggest that outcome-oriented individuals are less susceptible to the gamblers' fallacy: They report that the probabilities of getting a tail versus a head after four successive heads are equal. On closer inspection these individuals seem to mean by "equally likely" that either heads or tails *could* occur rather than that each has an equal probability of occurrence.

The apparent repeatability of trials is another problem variable that may encourage outcome-oriented responses. The weather, misfortune, and cab problems may have encouraged the prediction of individual trials because it was not easy to embed the event into a larger set of which it could be viewed as a sample (e.g., to embed a particular forecast of 70% chance of rain into the set "instances when the forecaster says 70% chance of rain"). In cases where trials were obviously repeatable (e.g., bone problem), however, some students still focused on single trials, even when the probability of a set of those trials was explicitly requested (as in the painted-die problem).

In summary, repeatability of trials and uniformity of elementary events are two problem variables that may influence the application of the outcome approach by inducing people to think of single trials or of causal factors, respectively. But even in the absence of such problem cues, some individuals will apply the outcome approach. This suggests that the outcome approach is not simply a problem-based phenomenon.

Outcome Approach Versus Personalist Interpretation

When requested, outcome-oriented individuals will attach numeric values to their predictions. It is clear from the bone and weather problems that these values are not akin to the frequentists' probabilities. As mentioned previously, the values associated with single-trial predictions appear to reflect degree of belief. In this respect the outcome approach is similar to the personalist interpretation; however, there are two important differences. First, personalist interpretations have been motivated by the desire to put subjective probabilities on a rational and scientific basis. Thus, among other requirements in these systems, subjective probabilities of repeated events should, over a long series of observed trials, closely approximate the actual frequencies of occurrence:

> If a person assesses the probability of a proposition being true as .7 and later finds that the proposition is false, that in itself does not invalidate the assessment. However, if a judge assigns .7 to 10,000 independent propositions, only 25 of which subsequently are found to be true, there is something wrong with these assessments. The attribute that they lack is called calibration. . . . Formally, a judge is calibrated if, over the long run, for all propositions assigned a given probability, the proportion that is true equals the probability assigned. (Lichtenstein, Fischhoff, & Phillips, 1981, pp. 306–307)

In explaining why people's probability judgments are poorly calibrated, Tversky and Kahneman (1982) suggested that because people do not naturally group events by their judged probability, they never have available the data that would permit them to make the necessary adjustments. Results on the weather problem suggest that, even when events are explicitly grouped in the appropriate way, these data are not used to evaluate the quality of probability estimates. The outcome-oriented individual appears uninterested in calibration as defined above but is interested instead in whether or not, on a particular occasion, a "correct" prediction can be made. If a nonpredicted result occurs, the prediction was wrong and the confidence value, if assigned, was too high.

The second and related difference between the outcome approach and the personalist interpretation is in the treatment of frequency information. Because a goal in a personalist interpretation is to be calibrated, the frequency of past occurrences of some event, when available, is used to formulate or adjust the initial probability. In the outcome approach, frequency data are not directly used to formulate confidence. It is especially clear in the cab, painted-die, and bone problems that frequency information, when considered, is first translated into a more qualitative belief from which a numeric confidence can be generated subsequently if it is requested. A similar two-stage process of generating subjective probabilities has been suggested by Adams and Adams (1961) and more recently by Koriat, Lichtenstein, and Fischhoff (1980).

> To assess one's confidence in the truth of a statement, one first arrives at a confidence judgment based on internal cues or "feelings of doubt". . . . The judgment is then transformed into a quantitative expression, such as a probability that the statement is correct. (Koriat et al., p. 108)

It should be added that the latter step of quantifying internal cues is probably not an essential component of the outcome approach outside the laboratory. It seems to be done, often grudgingly, only if a request for a percentage or probability is made. In the outcome approach, discriminating among small differences in the strength of these inner feelings is unnecessary. Given the goal of predicting the most likely outcome on a particular occasion, one needs to be aware only of which outcome is associated with the strongest inner feeling. It is difficult to imagine, in fact, how quantifying one's confidence could aid the decision-making demands of most day-to-day situations.

On the other hand, not being able to translate relevant quantitative information into belief strength is surely a handicap. Two possible reasons for this reluctance in using frequency data were mentioned previously: (a) that frequencies are viewed as an unstable source of evidence and (b) that they cannot be causally related to future events. Given only frequencies of past occurrence to predict future occurrence, it seems that the prediction would necessarily reflect the uncertainty represented in the distribution of past occurrences. But the outcome-oriented individual apparently has not accepted uncertainty as inherent in certain domains. They may even believe that someone who has mastered the mathematics of probability can predict the successive results of rolling a bone. As Student 9 responded, "If I were a math major, this would be easy."

Outcome-oriented individuals base predictions not on frequency data but rather on data that are deterministically linked to the event of interest. The importance of causality in making judgments under uncertainty has been demonstrated in a variety of contexts. Azjen (1977), Nisbett and Ross (1980), Tversky and Kahneman (1980), and others have demonstrated that distributional information is more likely to be incorporated into probability estimates if presented in a way that strongly implies a causal link between features related to the data and the event of interest. Similar to the event-sequence reordering observed in responses to the misfortune problem, people given biographies of deviants tend to reconstruct the information so that the plight of the "victim" can be viewed as an inevitable result of life events (Rosenhan, 1973). Also, the betting behavior of gamblers and the way in which they toss dice suggest that they believe they are controlling outcomes of chance events (Goffman, 1967).

If the outcome approach is a valid description of some novices' orientation to uncertainty, then the application of a causal rather than a black-box model to uncertainty seems the most profound difference between those novices and the probability expert and, therefore, perhaps the most important notion to address in instruction. As long as students believe that there is some way they can "know for sure" whether a specific hypothesis is correct, the better part of statistical logic and all of probability theory will evade them.

In this study, however, the preference for causal over stochastic models has been linked to the preference for predicting outcomes of single trials rather than sample results. As Kahneman and Tversky (1982) conjectured, "People generally prefer the singular mode, in which they take an 'inside view' of the causal system that most immediately produces the outcome, over an 'outside view' which relates the case at hand to a sampling schema" (p. 153). The fact that these two tendencies are not independent, but logically support one another, may explain in part why probability, as taught in the classroom, seems so foreign and difficult to master for many. Although the application of causal reasoning to stochastic processes may be the most salient difference between the outcome approach and formal theory, it may be more fruitful initially to have students focus on predicting sample results as opposed to single outcomes, thereby motivating a distributional schema.

Instruction in Misconception-Rich Domains

There is a more general implication of this and similar research into people's conceptions of probability and statistics, and that concerns the difficulties inherent in the teaching of domains in which people are known to hold strong prior conceptions (or *mis*conceptions) that are at odds with concepts central to the domain. In this regard students of probability and statistics and of physics face a similar problem. Beliefs students hold prior to instruction in physics and in probability and statistics interfere with learning the concepts introduced in the course. It is unfortunate that these misconceptions do not prevent many students from learning a host of the associated quantitative skills, because such skills can erroneously convince both teacher and student that the domain is being learned.

In physics education it has recently been advocated that instruction should encourage students to recognize and resolve conflicts between normative concepts and erroneous intuitions (McDermott, 1984). Several researchers (e.g., Clement, 1987; Hake, 1987; Minstrell, 1984) have demonstrated that physics instruction specifically designed to address various misconceptions can be effective. Their approach includes laboratory exercises designed to demonstrate counterintuitive results and promote student discussion, problems that require qualitative rather than quantitative solutions, and presentations that explicitly contrast normative with nonnormative physics concepts.

It has been suggested that a similar approach is necessary in the case of probability and statistics (Garfield & Ahlgren, 1988). In fact, the research of Kahneman and Tversky has already inspired curricula meant to alert students to their use of heuristics and to how these heuristics can lead them astray in judgments of uncertainty (Beyth-Myron & Dekel, 1983; Shaughnessy, 1981). My results suggest additional misconceptions that ought to be taken into account in the design both of probability curricula and of instruments meant to assess conceptual understanding. Given the current efforts to incorporate probability and statistics into the traditional K to 12 mathematics curriculum, such research ought to warn of the difficulties that some students will encounter, and serve as a guide to the conceptual areas that an effective curriculum will need to plumb.

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grants SED-8113323 and SED-8016567.

This article is based on a PhD thesis submitted to the University of Massachusetts.

I am grateful for the continuous support and reactions of Alexander Pollatsek and Amy Robinson and for the helpful comments on earlier drafts by Arnold Well, George Cobb, and Harry Schumer.

REFERENCES

Adams, J. K., & Adams, P. A. (1961). Realism of confidence judgments. *Psychological Review, 68,* 33-45.

Azjen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology, 35,* 303-314.

Beyth-Maron, R., & Dekel, S. (1983). A curriculum to improve thinking under uncertainty. *Instructional Science, 12,* 67-82.

Clement, J. (1987). Overcoming students' misconceptions in physics: The role of anchoring intuitions and analogical validity. In J. D. Novak (Ed.), *Proceedings of the Second International Seminar, Misconceptions and Educational Strategies in Science and Mathematics* (pp. 84-97). Ithaca, NY: Cornell University.

de Finetti, B. (1972). *Probability, induction and statistics.* New York: Wiley.

Falk, R. (1981). On coincidences. *Skeptical Inquirer, 6,* 18–31.

Garfield, J., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: Implications for research. *Journal for Research in Mathematics Education, 19,* 44–63.

Goffman, E. (1967). *Interaction ritual.* New York: Anchor.

Hake, R. R. (1987). Promoting student crossover to the Newtonian world. *American Journal of Physics, 55,* 878–884.

Howell, W. C., & Burnett, S. A. (1978). Uncertainty measurement: A cognitive taxonomy. *Organizational Behavior and Human Performance, 22,* 45–68.

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

Kahneman, D., & Tversky, A. (1972). On prediction and judgment [Whole issue]. *Oregon Institute Bulletin, 12*(4).

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

Kahneman, D., & Tversky, A. (1982). Variants of uncertainty. *Cognition, 11,* 143–157.

Konold, C. E. (1988). *Beliefs about equally likely vs. equally unlikely events* (Tech. Rep. No. 180). Amherst: University of Massachusetts, Scientific Reasoning Research Institute.

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

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1981). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Hueristics and biases* (pp. 306–334). New York: Cambridge University Press.

McDermott, L. C. (1984, July). Research on conceptual understanding in mechanics. *Physics Today,* pp. 24–32.

Minstrell, J. (1984). Teaching for the understanding of ideas: Forces on moving objects. In C. W. Anderson (Ed.), *Observing science classrooms: Perspectives from research and practice* (A.E.T.S. Yearbook XI). Columbus, OH: ERIC Center for Science, Mathematics and Environmental Education.

Monod, J. (1972). *Chance and necessity.* New York: Vintage.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90,* 339–363.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68,* 29–46.

Piaget, J., & Inhelder, B. (1975). *The origin of the idea of chance.* New York: Norton. (Original work published 1951)

Pollatsek, A., Konold, C., Well, A. D., & Lima, S. D. (1984). Beliefs underlying random sampling. *Memory & Cognition, 12,* 395–401.

Reichenbach, H. (1949). *The theory of probability.* Los Angeles: University of California Press.

Rosenhan, D. (1973). On being sane in insane places. *Science, 79,* 250–252.

Savage, L. J. (1954). *The foundation of statistics.* New York: Wiley.

Shaughnessy, J. M. (1981). Misconceptions of probability: From systematic errors to systematic experiments and decisions. In A. P. Shulte (Ed.), *Teaching statistics and probability* (pp. 90–99). Reston, VA: National Council of Teachers of Mathematics.

Todhunter, I. (1949). *A history of the mathematical theory of probability from the time of Pascal to that of LaPlace* (L. Lowell, Jr., P. Burrell, & H. D. Fishbein, Trans.). New York: Chelsea. (Original work published 1865)

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Review, 76,* 105–110.

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

Tversky, A., & Kahneman, D. (1980). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology* (pp. 49–72). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Tversky, A., & Kahneman, D. (1982). Judgment under uncertainty: Heuristics and biases. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 3–20). New York: Cambridge University Press.

→ Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunctive fallacy in probability judgment. *Psychological Review, 90,* 293–315.

von Mises, R. (1957). *Probability, statistics and truth.* London: Allen & Unwin.

Well, A. D., Pollatsek, A., & Konold, C. E. (1983). *Probability estimation and the use and neglect of base-rate information* (Tech. Rep. No. 66). Amherst: University of Massachusetts, Scientific Reasoning Research Institute.

APPENDIX

Problems: Interview 1

Weather Problem

What does it mean when a weather forecaster says that tomorrow there is a 70% chance of rain? What does the number, in this case the 70%, tell you? How do they arrive at a specific number?

Suppose the forecaster said that there was a 70% chance of rain tomorrow and, in fact, it didn't rain. What would you conclude about the statement that there was a 70% chance of rain?

Suppose you wanted to find out how good a particular forecaster's predictions were. You observed what happened on 10 days for which a 70% chance of rain had been reported. On 3 of those 10 days there was no rain. What would you conclude about the accuracy of this forecaster? If the forecaster had been perfectly accurate, what would have happened? What should have been predicted on the days it didn't rain? With what percentage chance?

Misfortune Problem

I know a person to whom all of the following things happened on the same day. First, his son "totaled" the family car and was seriously injured. Next, he was late for work and nearly got fired. In the afternoon he got food poisoning at a fast-food restaurant. Then in the evening he got word that his father had died. How would you account for all these things happening on the same day?

Bone Problem

I have here a bone that has six surfaces. I've written the letters *A* through *F,* one on each surface. [Student is handed the bone which is labeled *A, B, C,* and *D* on the surface around the long axis, and *E* and *F* on the two surfaces at the ends of the long axis.] If you were to roll that, which side do you think would most likely land upright? How likely is it that *x* will land upright? [Student is asked to roll the bone to see what happens.] What do you conclude about your prediction? What do you conclude having rolled the bone once? Would rolling the bone more times help you conclude which side is most likely to land upright?

[Student is asked to roll the bone as many times as desired.] What do you conclude having rolled the bone several times? How many times would you have to roll the bone before you were absolutely confident about which side is most likely to land upright?

One day I got ambitious and rolled the bone 1,000 times and recorded the results. This is what I got. [Subject is handed the list which showed $A = 50$, $B = 279$, $C = 244$, $D = 375$, $E = 52$, and $F = 0$.] What do you conclude looking at these? Would you be willing to conclude that *D* is more likely than *B*? That *B* is more likely than *C*? That *E* is more likely than *A*? If asked what the chance was of rolling a *D,* what would you say?

I'm going to ask you to roll the bone 10 times, but before you do, predict how many of each side you will get. How did you arrive at those specific values? [Student rolls the bone and records the results of each trial. After the 8th trial, the student is asked:] What is your best guess of what you will get on the next two rolls? [After the last trial, the student is asked:] How do you feel about your predictions? If you were going to roll the bone 10 more times, what would you predict that you would get?

Problems: Interview 2

Cab Problem

[Student is asked to read the cab problem aloud.] A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

1. 85% of the cabs in the city are Green, and 15% are Blue.
2. A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs, half of which were Blue and half of which were Green, the witness made correct identifications in 80% of the cases and erred in 20% of the cases.

What is the probability that the cab involved in the accident was Blue rather than Green? [After student gives a numerical response:] How did you arrive at that number?

Suppose the information in (1) were reversed such that 85% of the cabs in the city were Blue, and 15% were Green. The witness, as before, identified it as Blue and was 80% correct in the test situation. In that case, what would you say the probability was that the cab involved in the accident was Blue?

Bone 2 Problem

Last time you were asked which side of this bone you thought would most likely land upright. Do you remember which side you concluded? [The bone is held far enough away so that the labels cannot be read.] I'm going to ask you the same question again. And to give you something to base your answer on, I'll offer you any one of the following pieces of information. [Student is shown the list as the interviewer reads the items.]

1. A measure of surface area of each side.
2. The results of 100 rolls made by 16 people.
3. The results I got in 1,000 rolls.
4. A drawing of the bone showing the center of gravity.
5. The bone to look at.
6. The results on your last 10 rolls.

Which one would you like. Why did you choose that? If you could have a second piece of information, which would you choose? Why did you choose that? [Students are given both choices unless Item 4 has been picked. In that case, they are told that the drawing is not available and to pick another item. The estimate of surface area was in square inches: A = .028, B = *.078*, C = .065, D = .169, E = .018, and F = .031. The results of 100 rolls were: A = 7, B = 32, C = 21, D = 35, E = 5, and F = 0.] If you rolled the bone, which side do you think would most likely land upright? [Student is asked to predict the results of 10 trials; then the trials are conducted as in Interview 1.]

Painted-Die Problem

I have here a six-sided die. Suppose I told you that there was a possibility that it was loaded—that it had been altered so that one side was slightly more likely than the others to come up. Could you determine whether or not it was loaded? How? Would rolling it help you determine whether or not it was loaded? Suppose you rolled it 24 times and got the following results: [Student is shown the results as the interviewer reads them.] 1–5, 2–2, 3–8, 4–2, 5–4, 6–3. What would you conclude?

In fact, the die is not loaded. Suppose I painted five of the surfaces black and the other one white. If I rolled the painted die six times, would I be more likely to get six blacks or five blacks and one white? If I rolled it 60 times, how many times would you expect the white surface to come up? [This probe was originally worded, "On the average, how many times would you have to roll the die until you got a white?" After the third interview, it was changed to the present form, which was easier for students to understand.]

Obviously, I haven't painted the die. But I do have five black stones and one white one. [The stones were identically shaped pieces from a board game.] Suppose I put these in this cup and shook it really well. Then I reached in without looking and drew one out, wrote down the color, replaced it, shook it up again, and kept drawing like that. [This is demonstrated as it is explained.] Would that be the same as rolling the painted die? If I rolled the die several times and recorded what I got, and I drew stones and recorded those results, could you tell from looking at the results which I got from rolling the die and which from drawing stones? I'm going to draw six stones from the cup, but first ask you to predict what I'll get. [Stones are sampled, and before being shown the results of each trial, the student is asked to predict both the color that has been drawn and the probability that it is that color.]

Modeling Problem

You agreed that we could create a model of the painted die by drawing stones from a certain cup—that that would give comparable results. Would there be a similar way that we could make a model of the bone so that instead of rolling the bone, we could pick something out of a container and get the same kind of results?

[Student is given the following probes successively until a model is agreed on or the end of the list is reached:]

1. How about if we put six stones which have been labeled *A* through *F* in this cup and sampled from it as we did before?

2. Is there some container that I could fill with some number of lettered stones that would give results similar to rolling the bone?

3. Suppose we took the bone to a statistician and, however it is done, the following probabilities were calculated for each side: [Student is shown the list as the interviewer reads it.] *A* was 5 out of 100, or 5%; *B* was 29 out of 100, or 29%; *C*, 24; *D*, 37; *E*, 5; and *F*, 0. So, we took a big can and first put five of these stones which have been labeled *A* inside. [A large can and six small containers filled with labeled stones are placed in front of the student.] Then we took 29 *B*s, 24 *C*s, 37 *D*s, and 5 *E*s and put them in the container. Then we shook it up and sampled from it as before. Do you think that would give results comparable to rolling the bone?

4. Suppose we rolled the bone and, say, we got *B*. We took a stone labeled *B* and put it in the container. Then we rolled the bone again, and similarly, whatever we got, we put the appropriately labeled stone in the container, and we did that over and over. Would we reach a point when it would make no difference if we rolled the bone or drew from the container we had filled?

[When, and if, the student agreed on a model of the bone, the following questions were asked:] Suppose I rolled the bone 100 times and kept track of what I got. Then I drew 100 times from this can filled with the labeled stones. If I showed you the results from both, could you tell from looking at the results which I got from rolling the bone and which from drawing from the container? In the 100 trials with the bone and the container, do you think with one of those I'd be more likely than with the other to get no *E*s? Do you think I'd be more likely with one of those to get more *D*s in 100 trials than with the other?