1986

The effects of training on the quantification of uncertainty

Stephen P. Claxton-Oldfield
Wilfrid Laurier University

The Effects of Training on the Quantification of Uncertainty

by

Stephen P. Claxton-Oldfield

B.A. Queen's University, 1983

Thesis

Submitted to the Department of Psychology
in partial fulfillment of the requirements
for the Master of Arts degree
Wilfrid Laurier University
1986

ISBN 0-315-36048-8

Abstract

The use of intuitive heuristics has been put forward as an explanation for people's assessment of probabilities (Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1971, 1974). This phenomenon is seen as robust since "experts" (professional psychologists) make use of the same heuristics as "novices" (laypeople), despite having "had extensive training in statistics" (Tversky & Kahneman, 1974, p. 1130). However, replacing probability calculus with heuristics can lead to systematic errors and biases in probabilistic judgments. This study was designed to investigate the effects of statistical training on how people think about probabilistic judgments. Subjects' knowledge base of probabilistic concepts, as defined by the number of correct answers on the Probability Knowledge Questionnaire, was assessed prior to receiving (or not) a brief training session. Immediately following training, subjects completed a Probability Test which consisted of ten Tversky and Kahneman (e.g., 1974) problems. The findings suggested that the training served to make a statistical approach to the probability test problems more salient. It was also observed that "novices", with training, correctly solved a significantly higher proportion of test problems than did "experts" with no training. Finally, training served to increase subjects' judgmental accuracy as measured by confidence ratings. The training session, which was designed to sensitize subjects to some basic probabilistic concepts, was successful in reducing the use of heuristics.

Information on why people use heuristics and their robustness appears to point to a minimal probability knowledge base.

Acknowledgements

I wish to thank my thesis advisor Dr. Robert Gebotys for his guidance, patience, and smarts. I sincerely appreciate all the help he has given to me. I would also like to thank my thesis committee members Dr. Keith Horton and Dr. Angelo Santi for their constructive comments and valuable assistance. I must, of course, thank all my new friends in the Psychology Department for their support and for putting up with me over the past two years.

Finally, I would like to dedicate this thesis to my Mother for her continual encouragement and understanding.

Table of Contents

Tables

Figures

Every day, people are called upon to make decisions, and more specifically predictions, concerning the apparent chance of different events, for example, "how likely is it that it will rain tomorrow?", "what are the chances that a particular student will do well in graduate school?", and so forth. Some predictions are easy to make; others are more difficult. What makes predictions difficult is the existence of doubt or uncertainty. The uncertainty may stem from making predictions with incomplete information about some future state or event, that is, we do not have enough knowledge/the proper model (e.g., uncertainty about the future state of the market, or uncertainty about the weather) or because the information is neither reliable nor unbiased (e.g., information is out of date, inaccurate, or based on heresay). Uncertainty may also stem from making predictions when one is not sure which of several possible outcomes would be preferable (e.g., choosing among various job offers).

For some predictions, the uncertainty that people feel can be quantified and their beliefs concerning the chance of these events expressed in numerical form as objective probabilities. The objective probability of an event is defined as the limit of the relative frequency of the event, as the number of repetitions of the event increases indefinitely. Provided the event under consideration satisfies the characteristics of a random system (Gebotys, personal communication, 1986), people can objectively assess the probability of the event occurring so that uncertainty can be dealt with in as precise a fashion as possible. Three criteria must be satisfied for a system to be random. First, the event must be repeatable under essentially constant conditions (e.g., we can examine a large number of independent repetitions of the event, such as rolling a die or tossing a coin). Second, the relative frequency of the event is seen to converge or approach a limit as the number of events increases (i.e., is stable). For example, most people know that when the objective probability of an event occurring is .5, then half the time, on average, the event will occur. However, fewer people realize that the average may approach 50 percent only after the number of events becomes very large. That is, people do not understand relative frequency distributions. As an example, Fraser (1967) conducted an experiment in which he repeatedly tossed a die 12,800 times. He kept a continuous record so that at any stage in the tossing it was possible to read off the total number of 1's, 2's, ... obtained up until that time. As the number of tosses increased through the smaller values, the proportion of tosses yielding a given outcome fluctuated quite widely. For example, 50 repetitions yielded the following probabilities for p1, p2, ... p6:

p1=.280	p4=.180
p2=.260	p5=.040
p3=.120	p6=.120

As the number of tosses increased, these fluctuations became smaller and the proportions seemed to approach a limit. For example, 12,800 repetitions yielded the following probabilities for p1, p2, ... p6:

p1=.186	p4=.137
p2=.179	p5=.149
p3=.207	p6=.142

The third criterion of a random system is that the event must be random or nondeterministic. Although the relative frequency of an event approaches a limit as the number of repetitions of the system increases, there is no rule that can be formulated which will predict the occurrence of an individual event. Note also that randomness is not self-correcting. This is the gambler's fallacy and refers to the failure to appreciate the independence of some sequential events when the objective probabilities of these events do not depend upon previous events having occurred. For example, some roulette players feel that it is more probable that the ball will land on black after it has landed on three successive reds. Most games of chance (e.g., games with dice) satisfy these three criteria.

Objective probabilities may also be expressed as odds. The odds that an event with objective probability p will occur is defined as the ratio p:(1-p). Thus the odds that a 2 appears in the toss of a die (p2=.179 from Fraser's experiment) is .179:.821.

However, not all situations where people are faced with uncertainty satisfy the criteria of a random system. Many events are not repeatable. For example, we cannot repeatably put a particular student through graduate school in order to determine the objective probability that he/she will successfully complete the program. Another problem is that often the number of occurrences of an event is not large enough to approach a limit, or the conditions affecting the event are not controlled and held constant; in other words, the system is not random. In these situations, predictions may still be expressed in numerical form, but as subjective probabilities or odds. The term subjective probability refers to any estimate of the probability of an event when that event is obtained from a system which is not random.

Subjective and objective probability can also be considered from the perspective of Baye's Theorem. "Baye's Theorem is essentially an algebraic relationship by which prior probabilities are revised in view of additional data to obtain posterior probabilities" (Weber, 1973, p. 13). Objective prior probability refers to the objective probability of occurrence of an event in the future given that some previous event has occurred. For example, consider the conditional events tossing a coin and then drawing a single coloured ball from an urn. If the coin is tossed and it shows heads (H), then we go to an H urn containing 2 white balls and 1 black ball and draw 1 ball. If the coin shows tails (T), then we go to a T urn containing 1 white ball and 1 black ball and draw 1 ball. Given the coin shows H, we can objectively assess the probability for drawing a black ball or for drawing a white ball before we actually draw a single ball from the H urn. The realization of an outcome after repeatedly tossing the coin and drawing from the urns yields objective posterior probabilities. Note that prior probabilities do not always have to be objective. Baye's Theorem is used in many situations to revise prior probabilities as additional data (information) are obtained. The use of probability in many applications of Baye's Theorem has been criticized by statisticians who adhere to a classical (objective) interpretation of probability, that is, objective in the senses that prior probabilities are based on relative frequencies (Gebotys, personal communication, 1986).

Subjective probabilities represent a degree of personal belief about the probability of an event. Unfortunately, there is no objective probability against which one's beliefs can be compared. Tversky and Kahneman (1974) suggest that "the subjective assessment of probabilities resembles the subjective assessment of physical quantities such as distance and size" (p. 1124). Although determining the objective probabilities of uncertain events (when possible) is an accurate method of dealing with uncertainty, many people apparently do not understand or they fail to use this method when objective probabilities are readily computable (e.g., Kahneman & Tversky, 1972). According to Kahneman and Tversky (1972), "this is hardly surprising because many of the laws of chance are neither intuitively apparent, nor easy to apply. Less obvious, however, is the fact that the deviations of subjective from objective probability seem reliable, systematic, and difficult to eliminate" (p. 431). What determines people's beliefs about the probability of an event? Tversky and Kahneman (1974) argue that in a number of situations of judgmental uncertainty, including intuitive statistical judgments and categorical predictions (Kahneman & Tversky, 1972, 1973), people apparently rely on heuristics to assess probabilities and predict values (i.e., they replace the laws of chance with heuristics). A heuristic is a "rule-of-thumb" rapid form of reasoning that is assumed to yield reasonable (accurate) estimates, but quite often does not.

Some heuristics are highly economical; they can reduce the amount of time and effort normally required to make predictions under conditions of uncertainty. A research finding possibly related to the use of heuristics is that people are limited in their capacity to process information in short-term memory. This brief memory usually lasts under half a minute and the capacity to process information in short-term memory is usually limited to about seven items of information, give or take two (Miller, 1956). Some people write lists or diagram problems to cope with their limited processing capacity. Others prefer to work with a minimum of information (Kogan & Wallach, cited in DuBois, Alverson, & Staley, 1979) and are willing to move that information through the processing system quickly rather than thoroughly. Because time and effort are involved in acquiring information and because evidence indicates that the resources in short-term memory are limited, it may be more practical to use heuristics to make predictions under conditions of uncertainty.

People's reliance on heuristics has been regarded by some as a form of satisficing, that is, heuristics are used as cost-effective inferential shortcuts (e.g., Nisbett & Ross, 1980). However, heuristics often lead to systematic errors and biases in judgments. Heuristics "are less than perfectly correlated (if, indeed, at all) with the variables that actually determine the event's probability" (Bar-Hillel, 1982, p. 69).

Representativeness

One heuristic that is used to assign probability is that of representativeness (Kahneman & Tversky, 1972; Tversky & Kahneman, 1974). Kahneman and Tversky (1972) propose that people use the representativeness heuristic to judge, for example, how probable it is that a person is a member of a particular category. They infer that this is done by assessing the similarity, in terms of essential features (characteristics), of the person to the category based on the information available about the person and their knowledge of the category. The perceived similarity (indeed, it is a real increase in similarity) between the person and the category will increase with an increase in common features. They also propose that the representativeness heuristic may be used to judge how probable it is that an event originated from a given process. For example, could the sequence of coin tosses H H H H have occurred randomly? They suggest that this is done by assessing whether the event (H H H H) is seen to "reflect the properties of the uncertain process (randomness) by which it is generated" (Kahneman & Tversky, 1972, p. 434).

Although no general definition is available (it is an on-going area of research), representativeness can be assessed empirically, for example, by asking people to specify the level of similarity of an event(s) to a standard or to rate how much the event(s) reflects a given process. The concept of representativeness appears to be closely related to the concept of typicality (Rosch, 1975) According to Rosch (1975), some category members are seen as more typical of the category (better exemplars) than others. The most typical member of a category is called the prototype. The prototype does not have to be an actual exemplar but in some sense is an average of the most common properties of all exemplars in the category. It may be abstracted from experience or it may be empirically determined by rating the degree to which various category members are exemplary (i.e., typical) of the category (e.g., using a 10-point rating scale where a rating of 1 denotes a good example and a rating of 10 a poor example). Rosch and Mervis (1975) have shown that the prototypical member of a category is better learned, recalled, and recognized than other members of the category. The representativeness judgment, like typicality judgments, is also based primarily on the notion of similarity or resemblance (e.g., how much a person looks and acts like a member of the category).

For example, consider an individual who has been described as follows:

Steve is very shy and withdrawn, invariably helpful, but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail (Tversky & Kahneman, 1974, p. 1124).

Suppose you were asked to assess the probability that Steve is engaged in a particular occupation. Is he a farmer, salesman, airline pilot, librarian, or physician? Without adequate information about the frequencies with which the various occupations occur and personality characteristics of the people in these different occupations, one could use the representativeness heuristic to order these occupations from most to least likely. That is, one estimates the extent to which Steve is representative of, or similar to, the average or prototypical person in each of these occupations, based on our assessment of the members of each occupation. In the present case, the description of Steve is probably representative of, or similar to, our stereotype of a librarian and the subjective probability that Steve is a librarian is judged to be high. On the other hand, the description of Steve is not representative of, or similar to, our stereotype of a farmer, salesman, airline pilot, or physician and the subjective probability that Steve is engaged in any one of these occupations is judged to be low.

Base Rate Information

For another illustration of judgment by representativeness, consider an individual who has been described as follows:

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles (Kahneman & Tversky, 1973, p. 241).

Kahneman and Tversky (1973) told subjects (Israeli high school students) that the above description was randomly selected from a set of 100 descriptions of which 70 were of lawyers (engineers), while 30 were of engineers (lawyers). Subjects were asked to predict the probability that the above description and four other descriptions were of a lawyer (engineer): The probability that Jack is one of the 30 (70) engineers (lawyers) in the sample of 100 is ____%. The results indicated that subjects disregarded the prior probabilities (70-30) and apparently predicted that Jack was an engineer in both the high-engineer (70-30) and low-engineer (30-70) groups (producing essentially the same probability judgments) because the description of Jack was representative of, or similar to, the attributes that one stereotypically associates with engineers. Subjects did not compute objective probabilities for each description nor did they apply a Bayesian argument. Kahneman and Tversky (1973) concluded that since prior probabilities do not influence similarity under the above conditions, they are disregarded whenever the representativeness heuristic is used. It may be that the subjects' subjective stereotypes overrode the statistical properties, that is, subjects had both base rates and stereotypes and decided that the subjective stereotypes were more reliable or better predictors. Kahneman and Tversky (1973) inferred that subjects relied on comparing the features of the target, based on the description, with a set of features typically associated with their personal stereotypes of lawyers and engineers and predicted the outcome of which the target was most representative. However, Kahneman and Tversky's (1973) suggestion that subjects' judgments were controlled by the similarity of the description to their stereotypes is an inference based on the data. Kahneman and Tversky (1973) did not ask subjects what determined their beliefs (i.e., how they arrived at their probability estimates). Post-experimental interviews were not conducted to determine what strategies subjects had used to assign probabilities to the descriptions.

Carroll and Siegler (1977) suggest that perhaps Kahneman and Tversky's (1973) subjects viewed the category base rates as relevant but had difficulty drawing a direct implication from them. Given that there was not a direct correspondence between the 70-30 division in the population and how the sample of 5 target case descriptions in Kahneman and Tversky's (1973) study could be divided, subjects could not attempt to probability match (these figures did not afford a direct translation into a 7-3 partition for the 5 member sample) and, as a result, subjects decided to ignore the base rate information. If this was the case, it is possible that subjects did not know how to apply simple probability concepts. However, the ease with which a base rate may be applied to a prediction task may affect subjects' use of the information. This could be tested by having subjects predict the occupations of 10 individuals sampled from a population of 100 with a 70-30 base rate, thereby allowing for a direct translation into a 7-3 partition for the 10 member sample. There is some evidence that predictions are influenced by the translatability of the base rates (Carroll & Siegler, 1977, Experiment 2). However, the use of translatable base rates may be limited to situations in which uninformative personality descriptions (void of stereotypic content) are used (Carroll & Siegler, 1977, Experiment 3).

When subjects in the Kahneman and Tversky (1973) study were given no information whatsoever about the individual chosen at random from the sample (null description), they correctly used the base rates and made very accurate predictions. If subjects did not have some understanding of probability calculus, they would not be able to do a "better" job of assigning probabilities to the null description. This result suggests that when base rates are the only information available, subjects will draw on them as a basis for their inferences. Otherwise, subjects appear to be highly responsive to the personality descriptions and consequently do not use a Bayesian or classical application of probability. This finding lends some support to Kahneman and

Tversky's (1973) suggestion that subjects' judgments were controlled by the degree to which the descriptions appeared representative of their stereotypes. If we accept this interpretation, it is possible that the base rates in the first experiment were discrepant with subjects' expectations aroused by the target case descriptions. In other words, the personality characteristics of the target cases biased the subjects' predictions and their similarity-based judgments overrode the base rate information. Cohen (1981) has shown that occupational categories can bias memory toward consistent attributes, that is, people tend to remember characteristics when they confirm a stereotype. Thus, when subjects were asked to estimate the probability that Jack was one of the 30 (70) engineers in the sample, they remembered that he showed no interest in political and social issues and liked mathematical puzzles and this information overrode the base rate information. An experiment could be designed to determine whether subjects can correct these errors by telling them the hypothesized cause of the errors. For example, if subjects are told to try not to let their judgments be biased by their stereotypes because it can lead to errors, then they might use the base rate information. However, it may not be easy to convince subjects to put aside their stereotypes. An alternative would be to just clearly explain to subjects the rules of probability calculus.

When Kahneman and Tversky (1973) introduced a personality description which was uninformative with regard to profession, subjects disregarded the base rates and predicted that the target was equally likely to be a lawyer or engineer (0.5). The description read as follows:

Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues (Kahneman & Tversky, 1973, p. 242).

When there was no stereotypic content in the description, subjects may have found the information did not match their stereotypes of either occupation, and subsequently ignored the base rate information in favour of a 50-50 chance prediction.

In another experiment, Kahneman and Tversky (1972) gave subjects a list of nine areas of graduate specialization and asked subjects to estimate the percentage of students who are now enrolled in each. The subjects estimated that three times as many students were enrolled in the humanities and education than were enrolled in computer science. Then, an independent group of subjects read the following description:

Tom W is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and sympathy for other people and does not enjoy interacting with others. Self-centered, he none-the-less has a deep moral sense (Kahneman & Tversky, 1973, p. 238).

Subjects were asked to rank the nine areas in terms of how similar Tom W. is to the typical graduate student in each of the nine areas (a ranking of 1 indicated high similarity and a ranking of 10 low similarity). Subjects ranked Tom W. as much more similar to the typical graduate student of computer science (mean similarity= 2.1) than to the typical graduate student of humanities and education (mean similarity= 7.2) although they were surely aware of the fact that there are many more graduate students in the latter area. This result suggests that subjects made their predictions by judging the similarity of the description to their stereotype of the typical graduate student in each area. The data indicate that, for this problem at least, subjects used the representativeness heuristic.

A considerable amount of research suggests that prior probabilities or base rate information are ignored not only when making category membership predictions (e.g., Kahneman & Tversky, 1972, 1973), but also when making judgments such as behavioural predictions (e.g., Nisbett & Borgida, 1975) and causal attributions (e.g., Nisbett, Borgida, Crandall, & Reed, 1976). Nisbett and Borgida (1975) described to subjects the procedure sections of two conditions of two previously conducted psychology experiments: the high fear condition of a shock tolerance study (Nisbett & Schacter, 1966) and the emergency condition of a bystander intervention study (Darley & Latane, 1968). Some subjects were told about the actual distribution of behaviour in the experiments (base rates). The base rates indicated that most of the participants in the shock study tolerated a high intensity of shock and that most of the participants in the helping study helped only after a long delay or did not help at all. Subjects were asked to predict how target persons described as actual subjects in the experiments had behaved. A different group of subjects was asked to estimate the base rates in the two experiments. The results indicated that knowledge of the distribution of behaviour in the actual experiments did not influence subjects' predictions about the behaviour of target persons. Subjects did not use the base rate information to make their predictions; rather they predicted target persons would tolerate a moderate intensity of shock and would almost always help. The guesses made about the base rates by subjects lacking knowledge of the actual distribution of behaviour in the two experiments were highly similar to the subjects' predictions about the behaviour of the target persons. Thus, subjects ignored base rates for behaviour just as Kahneman and Tversky's (1973) subjects had ignored base rates for categorical predictions.

Nisbett and Borgida (1975) speculated that base rates, like other kinds of statistical data, are "remote, pallid, and abstract" whereas target case information is "vivid, salient and concrete." According to Nisbett and Borgida (1975), "vivid, salient and concrete" information is more likely to generate inferences because of the likelihood that such information will call up "scripts" involving similar information. The inference then proceeds along the lines of the previously existing script. "Remote, pallid and abstract" information is likely "less rich in potential connections to the associative network by which scripts can be reached" (Nisbett et al., 1976, p. 25). In other words, subsequent probability judgments are guided by the contents of scripts. Access to scripts may be more readily achieved if the information that calls to mind a particular script is "vivid, salient and concrete" (descriptive) rather than "remote, pallid and abstract" (statistical). An alternative explanation may be that people simply do not know when and, in many instances, how to use probability calculus when making predictions. Kahneman and Tversky's (1973) explanation for subjects' failure to be influenced by base rate information appears to center on the idea that people use heuristics in many situations.

Ajzen (1977) proposed that people ignore base rates in favour of descriptive information (e.g., personality descriptions) when base rates have no intuitive causal meaning, but that people will be influenced by base rates only "to the extent that they find it possible to incorporate the information within their intuitive theories of cause and effect" (Ajzen, 1977, p. 312). In other words, people will utilize information supplied by base rates only to the extent that they explain why a particular event is more likely to yield one outcome rather than another. Ajzen (1977) termed this the causality heuristic although it is a misuse of the term causality from the standpoint of statistics (Gebotys, personal communication, 1986). Ajzen (1977) proposes that people often rely on their intuitive understanding of the factors that they perceive to cause an outcome and they use this information for the purposes of prediction while disregarding the base rate information In one experiment (Experiment 1), the subjects' task was to predict the grade point average of 10 hypothetical students from 2 cues - one that was intuitively causal (IQ or study time) and one that was not (income or distance from campus). In addition, subjects were told that each cue had either a strong or weak statistical relation to the criterion as defined by the mean grade point average of students due to the influence of each of the cues The information concerning each cue was presented to subjects as summary data prior to the elicitation of predictions The results indicated that whether a cue provided intuitively causal or noncausal information had a profound effect on subjects' predictions (subjects were more influenced by causal and strong cues than by noncausal and weak cues). The perceived causal significance of the information had a greater influence on subjects predictions than did its statistical relation to the criterion (strong/weak). In fact, the causal cue was given greater weight than the noncausal cue even when the statistical relation between the causal cue and the criterion was weak while the relation between the noncausal cue and the criterion was strong. For example, when IQ was a weak predictor and distance was strong, the mean weight given to IQ was significantly higher than the mean weight given to distance It is possible that Ajzen's (1977) subjects ignored the information provided by the relation between the cue and the criterion simply because they did not believe it.

According to Ajzen (1977), the population base rates used by Kahneman and Tversky (1973) provided little, if any, intuitively causal information. That is, the proportion of engineers and lawyers in the sample "did not cause any member of the sample to become an engineer or a lawyer, nor did it provide information about any other factor that might be viewed as having a causal effect on a person's professional choice" (Ajzen, 1977, p. 304). However, the descriptions provided information (intuitive scripts) describing personality traits, interests, motivation, and ability which favoured either the lawyer, the engineer, or neither profession. Perhaps the information about the personality of the individual may be seen as intuitively causally related to profession, as Kahneman and Tversky (1973) suggest, because it was consistent with subjects' stereotypes of engineers and lawyers or, as Nisbett and Borgida (1975) suggest, because the information called up scripts involving similar information.

Ginosar and Trope (1980) tested the hypothesis that base rates will be used to the extent that the diagnostic usefulness (relevance) of the descriptive information is diminished or made less related to the judgment task. Using the engineer-lawyer problem (Kahneman & Tversky, 1973), they added to the descriptions a few characteristics with implications for membership in the occupational categories that conflicted with the other characteristics in the description and that should lower one's certainty that the target person is a lawyer (engineer). This was done by combining characteristics from Kahneman and Tversky's (1973) engineer description with characteristics from their lawyer description. A pretest indicated that these two kinds of characteristics had opposite implications for each profession. The inconsistent information condition read as follows:

Dan is a 45-year-old man. He is married and has four children. He is generally conservative, careful, ambitious, competitive, and argumentative. He is interested in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles (Ginosar & Trope, 1980, p. 233)

Ginosar and Trope (1980) also included the description constructed by Kahneman and Tversky to be representative of a prototypic engineer (consistent information condition) and the description constructed by Kahneman and Tversky to be uninformative or unrelated to the target person's profession (unrelated information condition). Subjects assessed the probability of a target person belonging to either of the outcome categories (lawyer-engineer) on the basis of category base rate level (70-30, 50-50, or 30-70) and the descriptive information about the target person. When Ginosar and Trope (1980) manipulated the information in the descriptions so that it was less related to the judgment task (inconsistent and unrelated information conditions), subjects ignored the information in the descriptions and used the base rate information. In the consistent information condition, the judged probabilities were similar under the various base rate levels and higher than they were for the other conditions. Thus, when the descriptive information was diagnostic of the outcome categories (i.e., representative of a prototypic engineer), the base rate information was disregarded. The finding that subjects incorporated base rate information into their judgments in the unrelated (nondiagnostic) information condition contradicts Kahneman and Tversky's (1973) finding; median estimates were 50% in both the high- and low-engineer groups under similar conditions. Still other studies have obtained intermediate results where the base rate was not ignored but rather diluted by nondiagnostic evidence (e.g., Manis, Dovalina, Avis, & Cardoze, 1980). Ginosar and Trope (1980) concluded that "where the actual description of a target person is ambiguous with regard to the categorization in question, people will abandon the simple, appealing strategy of exclusive reliance upon individuating information" (p. 240) and will incorporate base rates into their judgments. According to Ginosar and Trope (1980), it appears that subjects preferred to base their judgments on personality descriptions when the information in the descriptions was highly diagnostic but when the information was nondiagnostic, attention shifted to alternative, perhaps less preferred information (e.g., base rate information). This interpretation is similar to the one offered by Nisbett and Borgida (1975).

The results of studies by Shuford (1961) and Robinson (1964) showed that subjects' estimates of probabilities could be close to the true proportions under certain conditions (e.g., estimating the number of horizontal bars in matrices composed of horizontal and vertical bars in varying proportions). These findings suggest that when subjects make probability estimates based on a visual inspection of a total and well delineated population (well delineated in the sense that there are no stereotypes associated with population members), their estimates are accurate and unbiased. One obvious way in which these studies differed from the above experiments is that subjects' estimates were based on a visual display of the instances. A visual display of the base rates may have improved subjects' performance in the Kahneman and Tversky (1973) engineer-lawyer study. To test this notion, half of the subjects could be given a picture of 70 (30) hard hatted engineers and 30 (70) briefcase carrying lawyers along with the personality descriptions. It is possible that being able to see all the instances (100 engineers and lawyers by base rate level) and being able to have the display in front of them when making their predictions might cue subjects to use the base rates when making their judgments.

Another possible way to inform people about the applicability of base rates might be to give them the base rates in several different ways. Changing the base rates may be an important manipulation. Base rates have been changed between subjects in several studies (e.g., Kahneman & Tversky, 1973; Ginosar & Trope, 1980). Changing base rates within subjects might affect one's probability judgments. An alternative would be to just clearly explain to subjects the rules of probability calculus.

Sample Size

Another factor that people often ignore in judgments of representativeness is sample size. If we take random samples from a population, the larger the sample the more likely it is to be representative of the population from which it was taken. However, a number of studies have shown that people tend to overlook (or do not know) the principle that as the size of the sample increases the sample variance decreases. People tend to react to large and small samples in the same way, that is, as if any sample truly represents the population. According to Tversky and Kahneman (1974), people do not have valid intuitions corresponding to the impact of sample size on sampling variance. For example, Borgida and Nisbett (1977) found that subjects (first year psychology students) were quite willing to take a college course on the recommendation of one or two people without troubling themselves to seek out larger samples (e.g., mean course evaluations) when these larger samples were readily available. Indeed, when subjects were given mean course evaluations, this information had little impact on course choices whereas the brief face-to-face comments of 2 or 3 undergraduates had a substantial impact. This finding suggests that the opinion of a single, perhaps highly atypical individual may be taken as quite indicative or representative of opinions in general. Alternatively, if the person is seen as similar to the subject, the subject may decide that this one opinion is the best piece of information on which he/she should make a judgment. According to the representativeness heuristic, Tversky and Kahneman (1971) propose that "people believe samples to be very similar to one another and to the population from which they are drawn" (p. 106). Kahneman and Tversky argue that the role of sample size will be ignored owing to the application of a general heuristic, namely representativeness.

Tversky and Kahneman (1974) have shown that subjects fail to appreciate the impact of sample size on sampling variance even when it was emphasized in the formulation of the problem. Consider the following question posed by Tversky and Kahneman (1974):

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

For a period of one year, each hospital recorded the days on which more than 60 percent of the babies born were boys. Which hospital do you think recorded more such days?

a) the larger hospital
b) the smaller hospital
c) about the same (that is, within 5% of each other)
(Tversky & Kahneman, 1974, p. 1125).

The correct answer is that an extreme outcome is more likely to occur in the smaller hospital, because the outcome in a larger hospital is less likely to stray from 50 percent. Subjects apparently were not aware of the role of sampling variance. Most of the subjects (56%) judged the probability of obtaining more than 60 percent boys to be the same in both the large and small hospitals, "presumably because these events are described by the same statistic and are therefore equally representative of the general population" (Tversky & Kahneman, 1974, p. 1125). Perhaps if subjects were given information about the impact of sample size on sampling variance, they would be able to correctly apply the basic principles involved in the law of large numbers to this problem.

However, Tversky and Kahneman (1971) have shown that experienced research psychologists, who are supposed to have "had extensive training in statistics" (Tversky & Kahneman, 1974, p. 1130) also have a tendency to regard randomly drawn small samples as highly representative of the population. Tversky and Kahneman (1971) asked professional psychologists to respond to a questionnaire concerning research decisions at a meeting of the American Psychological Association. They found that experienced researchers are also prone to the same biases as untrained subjects. That is, although they avoided elementary errors (e.g., the gambler's fallacy), they displayed a tendency to put too much faith in the results of small samples and overestimated the replicability of such results. According to Tversky and Kahneman (1971), people's intuitions about random events are wrong in fundamental respects. Further, "these intuitions are shared by naive subjects and by trained scientists; and...are applied with unfortunate consequences in the course of scientific enquiry" (p. 105).

Tversky and Kahneman (1974) have also reported a similar insensitivity to sample size in subjects' judgments of posterior probability (see page 4). Consider the following problem:

Imagine an urn filled with balls, of which two thirds are of one colour and one third of another. One individual has drawn 5 balls from the urn, and found that 4 were red and 1 was white. Another individual has drawn 20 balls and found that 12 were red and 8 were white. Which of the two individuals should feel more confident that the urn contains two thirds red balls and one third white balls, rather than the opposite? What odds should each individual give? (Tversky & Kahneman, 1974, p. 1125).

The correct posterior odds for this problem are 8 to 1 for the 4:1 sample and 16 to 1 for the 12:8 sample. Tversky and Kahneman (1974) found that most people feel that the first sample provides much stronger evidence for the hypothesis that the urn is predominantly red, because the proportion of red balls is larger in the first than in the second sample. Thus, even though 4 out of 5 looks like better odds than 12 out of 20, the 12 out of 20 is the more reliable indicator. Large samples, randomly drawn from the population, have smaller variance (deviant sample percentages are more likely to occur in small samples than in large samples). Consequently, estimates derived from a large sample are more reliable.

Illusion of Validity

Judgments made on the basis of representativeness may also show an insensitivity to validity or predictive value. There are situations in which people appear insensitive to the relevance or quality of the information as a predictor of some outcome. For example, is the information accurate and fully reliable or out of date, inaccurate and based on heresay? In the Tom W. study by Kahneman and Tversky (1973), the more the description of Tom W. matched the subjects' stereotype of a graduate student of computer science, the higher was the subjects' confidence in their prediction. This gives the illusion of validity in the prediction even though subjects may have been aware of factors that should have limited the accuracy of their predictions and also their confidence. An independent group of subjects in the Tom W. study read the description of Tom W. (see page 14) along with the following information:

> The preceding personality sketch of Tom W. was written during Tom's senior year in high school by a psychologist, on the basis of projective tests. Tom W. is currently a graduate student (Kahneman & Tversky, 1973, p. 239).

Subjects in this group predicted which area of specialization Tom W. was likely to be studying and then evaluated the predictive accuracy of projective tests. Subjects expressed little faith in the predictive value of projective tests (which provided the basis for the description of Tom W.). However, using the representativeness heuristic, they overwhelmingly (95%) predicted that Tom W. was more likely to be studying computer science than humanities and education. Subjects obviously overlooked the point that people probably change rapidly following high school, so there was little reason to have trusted the description now that Tom W. is in graduate school; the description of his personality may no longer be accurate (Kahneman & Tversky, 1973).

Misconceptions about Regression

Regression to the mean is a phenomenon that is poorly understood by most people. Extreme outcomes, either high or low, are usually not followed by similar extreme outcomes. For example, suppose a group of students has written two equivalent versions of a test. If one selects ten students from among those who did best (worst) on one of the two versions, it will usually be found that their performance on the second version is not quite as high (low). Kahneman and Tversky (1973) suggest that "a major source of difficulty is that regression effects typically violate the intuition that the predicted outcome should be maximally representative of the input information" (p. 250). That is, people expect the initial high (low) performance of a student selected for consideration because of his test score to be representative of all future performances. People reasoning according to statistical principles should be reluctant to base inferences on samples yielding values that could be presumed to be extreme. However, these nonregressive intuitions persist in spite of the potential influence of random factors that can affect performance either favourably (e.g., student had studied the same examples as appeared on a test) or adversely (e.g., student had trouble sleeping the night before). Kahneman and Tversky (1973) believe that people do not expect regression in many situations in which it occurs.

However, some studies have shown that under some circumstances people can be induced to make more conservative or regressive inferences (e.g., Nisbett, Zukier, & Lemley, 1981; Zukier, 1982). Zukier (1982) asked subjects to estimate the grade point average of fellow students about whom they had read brief descriptions. The descriptions contained information predictive of an extreme grade point average only or, in addition, contained irrelevant information such as the fact that the student drives a Honda, always wears plaid shirts, and so on. It was found that when highly predictive information was diluted with irrelevant information, predictions become more regressive or conservative. In other words, information that is irrelevant may dilute the relation between a predictor and a criterion. This result is similar to the findings of Ginosar and Trope (1979): When information was nondiagnostic (irrelevant), subjects incorporated the base rate information into their judgments. Zukier found that irrelevant information, in the absence of base rates, diluted the information that was highly predictive of a high grade point average and, as a result, subjects' predictions became more regressive.

Misconceptions about Chance

Misconceptions about randomness have been shown to bias judgments by representativeness. People tend to view random events as unpredictable and fair, but at the same time they have quite well developed ideas about what chance events ought to look like. For example, a sequence of coin tosses that contains an obvious regularity such as H H T T H H T T is not considered representative of a random process whereas an irregular sequence of coin tosses such as H T H H T H T H is seen to reflect the randomness of the process (e.g., Kahneman & Tversky, 1972). When asked to judge which sequence is more likely to occur, people will erroneously pick the irregular sequence, because it looks random. According to Kahneman and Tversky (1972), in determining apparent randomness, people expect irregular sequences and local representativeness (i.e., even short sequences of coin tosses should include about the same number of heads and tails) and their presence contributes to the representativeness of randomness. However, as already noted, for a process to be considered truly random, it must satisfy the criteria of a random system (see page 2). In fact the regular sequence in the above example is statistically just as likely to occur. The probability of a head or tail on any single coin toss is 1/2. The probability of getting two heads in succession is (1/2) (1/2)=1/4. We can multiply the two probabilities in this case because they are statistically independent of each other. That is, whether a head or tail is thrown on any toss does not affect the probability of throwing a head or tail on the subsequent toss. Thus, the probability of getting two heads followed by a tail is (1/2) (1/2) (1/2), and of (H H T T H H T T) is (1/2) (1/2) (1/2) (1/2) (1/2) (1/2) (1/2) (1/2)=1/256. Further, the probability of the sequence (H T H H T H T H) is (1/2) (1/2) (1/2) (1/2) (1/2) (1/2) (1/2) (1/2)=1/256. Indeed, the probability of any sequence of eight heads and/or tails is 1/256, assuming a fair coin.

Much of the evidence regarding heuristics comes from the numerous demonstrations (many cited in this thesis) that, in a wide range of judgmental situations, people's probabilistic judgments frequently deviate from objectively correct solutions in a manner that is consistent with heuristics (e.g., Kahneman & Tversky, 1973). However, Olson (1976) has reported some apparent violations of the representativeness heuristic in people's judgments of probability. Olson (1976) investigated the specific nature of the observed biases in judgments reported by Kahneman and Tversky (1972, 1973, Tversky & Kahneman, 1971, 1974) in order to "determine the factors that make particular task and problem characteristics the salient ones with respect to which representativeness is judged" (p. 608). It was found that, in some contexts, these biases deviated sharply from the obvious predictions of the representativeness heuristic. Olson (1976, Experiment 2), studied people's inferences about a population proportion and a binomial (two outcome) sample. Subjects responded to the following problem:

> Consider two Quebec towns. Anglophones are a majority (55%) of the voters in Town A, but are a minority (35%) in Town B. There is an equal number of electoral ridings in each town. You have the voters' lists from all ridings in both towns. You randomly select a list from one riding, and observe that exactly 45% of the voters are Anglophones. What is your best guess - is the riding in Town A? or in Town B? (Olson, 1976, p. 602).

This problem is similar to Kahneman and Tversky's high-school program problem (1972, p. 433). According to Kahneman and Tversky (1972), subjects should choose the town with the Anglophone minority because it maintains the minority observed in the sample (such a result would be attributed to the representativeness heuristic). However, 70% of the subjects in Olson's (1976) study choose the town with the Anglophone majority. According to Olson, subjects responded to the absolute numbers in the problem, exhibiting a concrete thinking bias. "The sample percentage (45%) is potentially a subset of the larger (55%), but not of the smaller (35%), population percentage" (Olson, 1976, p. 603). This finding suggests that subjects may not be evaluating representativeness at all or subjects may be evaluating representativeness "with respect to (problem) characteristics whose salience was not anticipated by the theorist" (Olson, 1976, p. 606).

Availability

The representativeness heuristic is a quick though fallible method of making probabilistic judgments. It is the best studied of the heuristics. Alternatively, one may judge probability by assessing availability. According to the availability heuristic, one judges the probability of an event by the ease with which relevant instances or occurrences can be recalled or imagined.

Search Set Biases

For an illustration of judgment by availability, suppose you were asked to estimate the proportion of words (minimum of 4 letters) in the English language that begin with the letter r (e.g., road) versus words that have the letter r in the third position (e.g., carpet). Most people state that the letter r is more likely to occur in the first position (70% among Tversky & Kahneman's subjects, 1972). In fact, the letter r occurs more often as the third letter in English words (Tversky & Kahneman, 1972). Tversky and Kahneman (1972) believe that people use the availability heuristic to search for words that begin with an r or have an r in the third position. Because it is not possible to recall and count all instances of words with r in the first and third positions, subjects attempt to recall some instances and judge overall frequency by availability. Because it is easier to search for words by their first letter than by their third letter, words beginning with the letter r are judged as more numerous. With the focus of attention on the letter r, activation will spread from that letter to words beginning with it. This process tends to make words beginning with the letter r more available than other words. As a result, these words will be overrepresented in the sample that people take from memory to estimate the proportion of words beginning with the letter r in the population relative to those words with r as the third letter. That is, it is not possible to structurally prime words with the letter r in the third position and make them more available (Collins & Loftus, 1975). Alternatively, these words may not be easier to search, but easier to find (i.e., fewer noninstances are found with first letter cue than with third letter cue). This would constitute a retrieval bias due to the effectiveness of the cue.

Retrieval Biases

Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) conducted a series of experiments to study how people judge the frequency of death from various causes. In one experiment (Experiment 1), subjects were presented with 106 pairs of causes of death constructed from 41 different causes (for which good estimates of the true frequency exist). For each pair (e.g., stroke versus diabetes), subjects were to estimate which cause of death was more likely. Subjects were also to decide how many more times likely this cause of death was, as compared with the other cause of death given in the same pair. Subjects consistently overestimated low frequencies (e.g., floods were estimated to take more lives than asthma, although death from asthma is actually 9 X's more likely) and underestimated high frequencies (e.g., accidental deaths were judged about equal in frequency to death from all diseases, although death from all diseases is actually 15 X's more likely). Subjects also tended to overestimate the frequency of death due to all accidents, motor vehicle accidents, flood, tornado, and cancer and to underestimate the frequency of death due to smallpox vaccination, diabetes, lightning, tuberculosis, and asthma. Lichtenstein et al. (1978) proposed that the tendency to overestimate some causes of death was due to the unrepresentative coverage of these causes of death in the news, the possibility that subjects had more experience with these causes, and that these causes were easier to imagine and more memorable. All factors that would tend to make these causes of death easier to retrieve and, according to the availability heuristic (Kahneman & Tversky, 1973), appear more likely than an equally frequent cause of death with less easily retrievable instances. According to Nisbett and Ross (1980), the salience or vividness of information can bias frequency judgments because it is more likely to attract and hold our attention and be more available when judgments are made.

The availability heuristic can also account for temporary shifts in subjective probability. For example, if you have just witnessed an automobile accident, your subjective estimate of the probability of this kind of event rises dramatically, although only for a short time (Tversky & Kahneman, 1974). One possible reason for this phenomenon is that recently experienced events are easily retrievable and are thus highly available. Also, these events establish a context in which we tend to retrieve information (call up scripts) about events of a similar nature (Nisbett & Ross, 1980). As a result, people tend to think they have a higher frequency of occurrence (overestimate their frequency) than they actually do.

Representativeness and availability are not the only heuristics people use. Unlike the more idiosyncratic heuristics that can be applied to one's job and hobbies, however, representativeness and availability can be applied across a wide variety of situations. People are not aware of the errors that often result from using heuristics. The basic explanation for people's errors appears to center on the idea that people do not understand probability calculus. Nisbett, Krantz, Jepson, and Kunda (1983) argue that "there is good reason to believe that people possess statistical heuristics ... intuitive, rule-of-thumb inferential procedures that resemble formal statistical procedures" (p. 345). However, it has been shown that in a wide range of judgmental situations that require such procedures, people often do not reason statistically and they often do not do so even if they have had formal training in statistics (e.g., Tversky & Kahneman, 1971).

Pilot Study

The purpose of conducting a pilot study was to investigate how people attempt to solve probability problems (i.e., problems for which objective probabilities are computable). When asked to predict the apparent likelihood of some event happening, people often express their judgments in terms of subjective probabilities. Most people apparently do not understand the distinction between objective and subjective probability and often fail to apply the proper normative principles in order to determine an event's probability. This is likely due to a lack of knowledge or absence of training in probability or statistics.

One purpose of this pilot study was to determine whether or not people can solve some probability problems as a function of their mathematical background (e.g., number of high school and university mathematics or statistics courses taken). Do people who have course training training in traditional statistics courses understand probabilistic concepts better than people who have had no formal training?

Subjects were asked to estimate the probability that certain individuals described in brief personality descriptions were engaged in a particular occupation. The occupations and prior probabilities were given to subjects before they read the description. Kahneman and Tversky (e.g., 1973) found that when they gave subjects base rate information (prior probabilities) regarding occupational membership (e.g., the following personality description was randomly selected from a set of 100 descriptions of which 70 (30) are of engineers and 30 (70) are of lawyers), people ignored the prior probabilities and apparently judged the probability that the individual described was an engineer (lawyer) on the basis of how much the description matched their stereotype of an engineer (lawyer). Kahneman and Tversky (1973) proposed that subjects used what they have termed the representativeness heuristic (which is a similarity-based strategy) to solve these problems.

A further purpose was to determine if people actually use representativeness (provided they fail to calculate the objective probabilities) or, in fact, use other heuristics. Kahneman and Tversky (1973) never asked subjects how they arrived at their judgments, but rather inferred subjects were using representativeness. Finally, some subjects in the pilot study were presented with a visual display of the base rate information (prior probabilities) to see whether this information cued them to use the base rates when making their judgments.

Method

Subjects

Subjects were 16 undergraduate volunteers: 7 females and 9 males. Subjects ranged in age from 19 to 47. Eleven first year students and 5 third year students from Wilfrid Laurier University served as subjects. Five subjects had no university mathematics and 4 of the 5 had no grade 13 mathematics. These five students were in their first year of university. The 11 other students had taken, on average, 2 mathematics or research methods courses at the university level. Of these 11, only two had no grade 13 mathematics.

Procedure

Subjects were informed they were participating in a study of people's intuitions about chance. The study was conducted in an informal interview format. All interviews were tape-recorded and basic demographic information was sought from all subjects before the study began (e.g., age, number of mathematics and/or statistics courses taken, etc.). In the first part of the study, all subjects were presented with a series of probability statements and were asked to describe, in their own words, what was meant by each of the statements. Some statements expressed objective probabilities, for example, the probability of an odd number appearing in the single toss of a fair coin is .5; others expressed probabilities that were subjective, for example, the probability of a particular student graduating is .9. Subjects read each statement aloud.

In the second part of the study, subjects were asked to solve a series of probability problems and to rate their confidence in their solutions. Subjects rated their confidence on a scale of 1 to 5 with 1 being very confident and 5 being not very confident. The six problems were presented in a randomized order for each subject. For each problem, objective probabilities were computable. Two of the problems were ball and urn type problems, for example, A coloured ball was randomly drawn from an urn containing a sample of 100 balls of which 70 were blue and 30 were red. What is the probability of drawing one of the blue balls? Two of the problems involved the concepts of combinations and permutations, for example, determine the probability that at least two people in a group of 23 randomly selected people share the same birthday. Finally, two Kahneman and Tversky (1974) problems were included. Subjects were to estimate the probability that a particular individual described in a brief personality description was engaged in a particular occupation. They were told that the description was randomly selected from a set of 100 descriptions of which 70 were of lawyers and 30 were of engineers. After subjects had attempted to solve each problem, the experimenter asked them a number of questions about how they had arrived at their solutions. The probability statements and problems appear in Appendix A.

Five subjects were presented with a visual display of the base rate information. Some subjects were shown a picture of 100 balls, 70 of which were blue and 30 of which were red. This information was also presented in a bar graph. Similar information was presented along with one of the Kahneman and Tversky problems. Subjects were shown 70 yellow squares (denoting lawyers) and 30 blue squares (denoting engineers) with a similarly coded bar graph.

Subjects were allowed to work at their own speed. The tape-recorder was turned off while subjects worked toward a solution for each problem and turned on for questions after subjects had made an estimate and had rated their confidence. For example, subjects were asked to indicate how they attempted to solve each problem (e.g., what specific information in the problems did they find most useful). The probability statements and problems were presented to subjects in a 7-page booklet.

Results

Probability Statements. Subjects' interpretations of the probability statements were vague. No subject made the distinction between subjective and objective probabilities. A probability of .3 was interpreted as meaning a "slight chance" or "relatively low chance" that the event (e.g., rain tomorrow) will happen. Four subjects interpreted it as a 3 in 10 chance of rain while others interpreted it as a 70% chance that it would not rain tomorrow. The modal interpretation for these statements was "poor chance" or "very unlikely" for probabilities less than 50% and "good chance" or "very likely" for probabilities greater than 50%. When the probability was .5 (odd number appearing on a single toss of a fair die), most subjects stated that there was a 50/50 chance that an odd number would appear. Only one subject mentioned relative frequency distributions and that in any short sequence of die tosses there was the possibility of tossing more odds than evens. This subject had university and grade 13 training in mathematics. When the notion of sample variance and sample size was explained to subjects (as an afterthought), all but two indicated they believed the proportion of odds and evens was more likely to approach .5 after 1000 tosses than after 10 tosses.

Ball and Urn Problems. Of the 16 subjects, only two were in error. Both of these subjects missed one of the ball and urn problems but were correct on the other. These two subjects had no university mathematics. All other subjects correctly used the relative frequencies (base rates) to solve these problems. Mean confidence ratings for the Blue Ball and Red Ball problems were 1.7 and 2.0, respectively.

Birthday and Married Couples Problems. No subject correctly solved either of these probability problems. In fact, most subjects admitted to guessing although some did attempt calculations. The correct objective probability for the Birthday problem is slightly greater than .5 or 50%. Most of the estimates (subjective probabilities) were less than 10% (11 subjects). When asked whether their estimates seemed reasonable, most subjects replied yes and some thought their estimate should have been even smaller. For the Married Couples problem, estimates were higher than for the Birthday problem although the objective probability that two married couples are selected is .03. Again, subjects felt their estimates were reasonable for this particular problem. Not surprisingly, subjects' confidence ratings for these problems reflected their uncertainty (mean confidence ratings of 3.5 and 3.6, respectively)

Kahneman and Tversky Problems. Seven subjects made incorrect estimates for the Kahneman and Tversky problem when they were asked to estimate the probability that the person described in the paragraph was a lawyer. Of these 7, 3 had no university or grade 13 mathematics. Six of these 7 subjects relied exclusively on the information contained in the personality description while disregarding the base rate information. Each of the six subjects felt that the statement "he shows no interest in social and political issues" decreased the probability that he was a lawyer, for example, "lawyers are more likely to have these interests," "most lawyers go into politics," etc. The same subjects were also influenced by the target person's hobbies which included "home carpentry" and "mathematical puzzles." These hobbies were "more associated with engineers" and "increased the probability he is an engineer rather than a lawyer." The data indicate that, for this problem, some subjects used the representativeness heuristic.

Only one subject attempted to use both sources of information (base rates and personality description). This subject realized that 70% of the sample consisted of lawyers but felt that because of the personality description there was "more of a chance he was an engineer." This subject had no university or grade 13 training in mathematics.

Of these 7 subjects, only 1 subject incorrectly estimated the probability of 1 of the ball and urn problems. Subjects knew how to use the base rate information for the simple probability problems; subjects knew how to estimate objective probability in the absence of a personality description.

Three of the nine subjects who correctly estimated the probability of the target person being a lawyer also indicated that the description "sounded more like an engineer" but added that the personality information "had nothing to do with probability," and "could not change the fact 30% are going to be engineers and 70% are going to be lawyers." Only 1 subject believed the description was there to "trick" him. All 9 subjects relied exclusively on the base rates and "did not worry about the description," for example, "he could be a lawyer and still like mathematical puzzles." These 9 subjects had university and/or grade 13 training in mathematics

Visual Base Rate Information. The five subjects who received the visual base rate information reported they did not use the visual displays when making their judgments for the two problems and they did not perform differently from those subjects who did not receive the visual displays.

Discussion

The results of the pilot study suggest that many subjects have not learned the proper application of some of the probabilistic concepts tested. Although most of the subjects had some training in statistics, they had difficulty applying the proper concepts to some of the problems. For example, most subjects admitted to guessing at their solutions to the Birthday and Married Couples problems. This finding suggests that subjects did not understand the concept tested or the concept had not been learned. Although training seemed to help on some problems (e.g., ball and urn problems), it did not carry-over to all problems. In fact, of the subjects who had no training, only two erroneously solved the ball and urn problems. It appears that subjects, even if they have not had formal training in statistics, understood the applicability of base rates for the ball and urn problems. That is, people seem to have an ability to use base rates for selected kinds of problems. On the other hand, having had some formal training in statistics was not sufficient to guarantee the proper application of probability calculus to all problems.

Training

The failure to use probabilistic concepts can affect people's inferences in everyday life, that is, people do not realize the limits of heuristics when compared to more objective methods. It may be possible to sensitize people to the basic concepts of probability and statistics and thus reduce their reliance on judgmental heuristics. The concepts are highly trainable and once mastered, the application of these concepts may become very rapid and even automatic (Nisbett et al., 1983). No doubt, many of the errors people exhibit in their judgments reflect a minimal probability knowledge base. For people who have received formal training, there are likely large individual differences in education and in practice that affect reasoning. For people who have not received formal training, a rough intuitive understanding of some statistical concepts may not be absent from their judgmental repertoire (e.g., Nisbett et al., 1983).

Fong, Krantz, and Nisbett (1986, Experiment 4) suggest that statistical training can have an effect on people's reasoning about everyday problems. Males in an introductory statistics course were asked to participate in a telephone survey of opinions about sports. The survey took place during the first two weeks of term and near the end of term. Subjects were asked five questions for which the application of statistical concepts such as the law of large numbers and the regression principle were relevant. For example, why are batting averages of .450 common during the first 2 weeks of the baseball season but unheard of as a season average? Subjects who rated themselves as having little or no knowledge of sports were not used in the experiment. Subjects' responses to the questions were tape-recorded and coded for the presence of statistical reasoning and for whether a statistical response was a good one. Fong et al found that near the end of term there was an increase in the percentage of statistical answers for this problem (50% of the answers were statistical at the beginning of term while 70% were statistical at the end of term). This was attributed to the single course in statistics. However, it should be noted that the meaning of the word average in the "batting average" question implies that a player with a seasonal average of .450 must have some scores higher than this. Mathematics does not determine whether these high scores, or weeks with high average scores, should occur early, late, or mid-season. The application of regression principles to this problem provides no explanation at all. In this study, there could have been a better selection of problems for testing the effects of statistical training. Fong et al. offer no explanation for why the statistics course did not have any effect on two of the five questions asked.

Fong, Krantz, and Nisbett (1986) examined the effects of a brief (25-minute) training procedure on how people think about everyday problems (Experiments 1 and 2). Some subjects received a four-page description of the concept of sampling and the law of large numbers and were shown it in operation by drawing various sized samples of gumballs from a vase (rule training). Some subjects received three example problems with an answer following each problem that provided an analysis of it in terms of the law of large numbers (examples training). Other subjects received rule training followed by examples training (full training) or no training. After training, subjects were given a test consisting of 18 problems divided into three major types (objective, subjective and probabilistic). For example, subjects had to draw conclusions about the outcomes of athletic events (objective), to decide which college a high-school senior should choose based on his own and his friends' reactions to the colleges (subjective), and objective problems with a probabilistic component (e.g., an explicitly random selection procedure). These problems were not the kind of problems that had been investigated by Kahneman and Tversky (e.g., 1973), that is, problems to which subjects apparently applied heuristics. A sample of 20 test booklets was coded by four coders with exact agreement on 86% of the problems.

The remaining 327 test booklets were coded by one coder using a 3-point system. Responses were coded as entirely deterministic (i.e., no mention of sample size and variability), poor statistical (i.e., incomplete or incorrect application of the law of large numbers), or good statistical (i.e., some form of the law of large numbers was used). Training effects, with respect to the frequency and quality of subject's statistical responses, were found across all three problem types. It should be noted that a poor (incorrect) statistical response was used as evidence that training increased the frequency of a statistical approach. However, the training enhanced statistical thinking about subjective, objective, and probabilistic problems of an everyday nature whether the training was only in the objective domain (Experiment 1) or in one of the three problem domains (Experiment 2) with the training domain varied as a between-subjects variable. For example, "subjects taught examples in one domain learned no more about how to solve problems in that domain than they did about how to solve problems in other domains" (Fong et al., 1986, p. 275). That is, there were no domain-specific effects of training. Fong et al. (1986) comment that:

> A qualification that must be placed on the present results is that the effects at least of relatively brief training sessions may be limited to problems for which some untrained subjects are able to give statistical answers. Many previous demonstrations of people's difficulties with statistical principles are based on problems to which no subjects, or almost no subjects, apply statistical reasoning (e.g., Hamill, Wilson, & Nisbett, 1980; Kahneman & Tversky, 1972, 1973; Tversky & Kahneman, 1983). Quite deliberately, we avoided such difficult problems in the present investigations (Fong et al., 1986, p. 281).

A quicker method has been suggested by Nisbett, Krantz, Jepson, and Fong (1982). They proposed the development of statistical heuristics that people can use in everyday inference, such as "You can always explain away the exceptions" which is a heuristic to remember the concept of variance or "Think about evidence as if it were a sample, and reflect about sample size." They hypothesized that people can be taught to be more sensitive to considerations of sample size and sampling variance. The notion of sample size, however, is not incorporated into people's intuitive heuristics. For example, representativeness does not appear to be influenced by sample size (i.e., size of the hospital) in the Kahneman and Tversky maternity-ward problem. In this case, the needed concept is that large samples are more likely to be representative of the population from which they are taken than are small samples. Statistical heuristics encourage people to think of information in terms of data properties and people do not have to be statisticians to use them. However, using rough intuitive statistical principles may not be dissimilar to guessing or using nonstatistical heuristics.

Another solution is to get people to turn their problems over to computers or experts. Computers do a reliable job of weighting and combining probabilistic information. However, this solution is not very practical. Using a computer to solve probability problems without knowledge of probability calculus is unreasonable from the point of view of many experts. Novices make mistakes applying the computer packages just as they do in many of the experiments discussed earlier.

Previous research in human inference has viewed the quantification of uncertainty as a "complex" problem where individuals apply heuristics (i.e., representativeness and availability) in order to simplify the inference process. Normative statistical methods are used to determine correct solutions to these problems and heuristics are inferred from incorrect answers. This phenomenon is seen as robust since experienced researchers (e.g., PhDs in psychology who have had training in statistics) make use of the same heuristics and are prone to the same errors as laypeople. According to Tversky and Kahneman (1971), statistical training does not seem to guarantee the use of probability calculus or eliminate errors of inference It may be that the training or knowledge base was not adequate (see page 44).

Other authors who have investigated errors of inference have tried to eradicate heuristics with limited success. Henrion and Hogarth (cited in Fischhoff, 1982) suggest that "improvements" in the quantification of uncertainty can be made by recalibrating responses, that is, if evidence in a situation leads to higher/lower than normative estimates, they encourage subjects to lower/raise their probabilities. Others (e.g., Fischhoff & Beyth, 1975) have urged subjects to work harder on problems and used contemporary events or examples. All the above techniques were employed to investigate the process of inference with the hope of gaining more insight into how people conceptualize uncertainty. This research strategy is one in which the system inputs are changed and if a change does or does not occur in the system, information is gleaned on how the system functions.

The present research addresses the following two issues in human inference:

1) Why do people use heuristics?

2) How robust are the effects of heuristics?

The human inference process can be conceptualized as the product of an interaction between person and task. People have the cognitive ability to master the concepts necessary to solve problems. However, they require a minimal probability knowledge base or else they will use heuristics. A minimal probability knowledge base can be "built" by giving people simple explanations of some basic probabilistic concepts that will be helpful in solving probability problems from a normative point of view. The pilot study clearly indicated that people had little knowledge of the normative definition of probability calculus. It is interesting to note that people have difficulty with probabilistic concepts in some situations without stereotypic or descriptive information. It may be that people use heuristics because they do not understand the questions. The first step in the solution of these inference problems is an understanding of the question. The second step is an application of the normative rules of probability and statistics (remember, a problem is correct or incorrect from this normative point of view) to obtain the answer. The person-task system is seen as having the cognitive skills necessary to solve problems but due to the lack of a probability knowledge base continues to use heuristics.

The present training strategy addresses the issue of task difficulty by giving clear instructions to people to ensure an understanding of the problem and consequently discourage second guessing. The probabilistic concepts were demonstrated and feedback was given on the errors made on the task questions. The task was decomposed to simple units of understanding where the concepts necessary were made explicit. Kahneman and Tversky (1979) have suggested in a theoretical context (i.e., no data) a global approach to the study of heuristics:

> The adoption of an external approach that treats the specific problem as one of many would help overcome this bias (Kahneman & Tversky, 1979, p. 314).

One goal of the present training study was to understand why people use nonprobabilistic heuristics. The Fong et al. (1986) training studies did not address the issue of why people use the representativeness and availability heuristics. They deliberately avoided using problems of the type that had been investigated by Kahneman and Tversky (e.g., 1972, 1973). In answering such problems, Kahneman and Tversky's subjects typically relied on heuristics to make probability judgments. To evaluate the effectiveness of training on people's use of representativeness and availability and to gain a better understanding of why people use heuristics, the problems from which these heuristics have been inferred have to be studied. The present study differs from the Fong et al. studies in other respects. For example, several probabilistic concepts were trained in the present study whereas only one concept was trained in the Fong et al. studies. The effects of training in the present study were investigated from a normative standpoint emphasizing objectively correct solutions to problems. In the Fong et al. studies, a coding system was used to distinguish between subjects' open-ended answers on the basis of whether or not a statistical response was "good" or "poor" (such a coding system likely runs into borderline cases). It is felt that the approach in the present study, based on sensitizing people to normative probabilistic concepts and on testing these concepts with Kahneman and Tversky problems, will be the first step in a fruitful investigation of this system.

Normative probabilistic solutions to these problems are provided by experts (i.e., people who understand probabilistic concepts). These people have an understanding of how to apply probability calculus. It is interesting to note that statistically sophisticated researchers, defined by Tversky and Kahneman (1974), that is, professional psychologists, use heuristics. This result is used as evidence for the robustness of heuristics. This thesis investigates whether statistically trained researchers (i.e., PhDs in psychology, as studied by Tversky & Kahneman, 1971) understand normative probabilistic concepts. If they do and still rely on heuristics to assign probabilities to all problems, the robustness of the system is substantiated: if they do not, then information on why people use heuristics and their robustness points to a minimal probability knowledge base. This hypothesis is tested by comparing the results of training versus no training conditions. If a 15-minute training session can reduce the effects of a robust phenomenon (i.e., reliance on heuristics), the effect is in fact weak and the problem can be conceptualized as one of understanding. Would one be surprised if a member of the public were given an auditor's report of a company and, when asked to assess the profitability of the company, give some rather atypical answer? It is suggested here that in both the example above and in human inference in general an understanding of normative rules is a reasonable approach to the problem. Previous approaches have concentrated on changing descriptive aspects of the problem (e.g., Kahneman & Tversky's, 1972, engineer-lawyer problem) and seeing how these manipulations affect probability estimates. It should be noted that in many instances, when people are asked to solve probability problems without stereotypic information, heuristics are still used. In other words, the descriptive approach permits a reasonable investigation of human inference at one level, whereas the argument presented in this paper approaches the problem from both the probability calculus and descriptive points ~~of~~ view.

Kahneman and Tversky (1979) speak of correcting heuristics through a combination of calibration and statistical maxim techniques. They conclude their article with:

> The analysis of human judgment shows that many biases of intuition stem from the tendency to give little weight to certain types of information, for example, the base-rate frequency of outcomes and their probability. The strategy of debiasing in their paper attempts to elicit from the expert relevant information that he would normally neglect and to help him integrate this information with his intuitive impressions in a manner that respects the basic principles of statistical prediction (Kahneman & Tversky, 1979, p. 327).

This paper is in agreement with the above important aims and advocates an approach to finding the source(s) and limits of heuristics that concentrates on sensitizing people to the concepts and thus building a minimal probability knowledge base.

A minimal knowledge base of probability calculus can sensitize people to manipulate information using probabilistic concepts. That is, when making probabilistic judgments, people can apply these concepts and avoid inferential errors. Sensitizing subjects to the concepts is the first step in addressing this issue. Moreover, training may improve the accuracy of subjects' probabilistic judgments. Oskamp (1965) examined the relationship between level of confidence and accuracy in a study of clinical judgment. Oskamp found that subjects' confidence increased as they were given more information for making their judgments. However, there was no significant increase in accuracy with increasing information. Other findings in research on people's judgmental expertise suggest that people are overconfident about their true predictive abilities (e.g., Hoch, 1985). Overconfidence may arise as a result of biases in evaluating the information on which predictions are to be based (Hoch, 1985). If overconfidence is similar to other cognitive biases, for example, the biases that are found in people's intuitive judgments of probability (biases to which heuristics lead), then training on normative probability calculus may improve people's accuracy of judgments by making them aware that they may be incorrect when they use heuristics.

The proposed training strategy will serve to make a statistical approach more salient. On the other hand, having completed a number of statistics courses is not sufficient to guarantee the proper application of probability calculus (Tversky & Kahneman, 1971). For example, Nahinsky and Ash (1985) found that, when asked to judge the likelihood of one trait (A) given another trait (B), as well as the likelihood of Trait B given Trait A, a group of 12 psychology graduate students, "trained in basic probability," applied judgmental strategies that did not include the application of probabilistic concepts. For convenience in the present study, experienced researchers, as defined by Tversky and Kahneman (1971), that is, people who have had extensive training in statistics, were classified as experts. Laypeople, as defined by Tversky and Kahneman (1971), that is, people who have had little or no training in statistics, were classified as novices. This implicit classification scheme was called the TVERSKY and KAHNEMAN classification factor. Another way to classify individuals rather than on the basis of their background in statistical training would be to develop a measure of the individual's probability knowledge base. To this end, a

Probability Knowledge Questionnaire was developed by Gebotys and Claxton-Oldfield (1986). The questionnaire, which attempts to measure a person's knowledge of probabilistic concepts, consists of ten probability problems (see Appendix B). The problems involve the probabilistic concepts of sample size, randomization, base rates, combinations, and correlation.

Finally, subjects were paid if their answers to the Probability Knowledge Questionnaire and Probability Test problems were correct. The potential of making money in return for correct answers was incorporated primarily to interest subjects in participating in the study. Kahneman and Tversky (1972) paid subjects $1.00 for participating in an experiment concerning sampling distributions and an additional $1.00 if subject's answers to one of three problems (randomly selected after completion of the task) was correct.

Overview

A total of 80 subjects took part in the experiment All subjects were cross-classified according to two criteria. Subjects were selected and classified according to the TVERSKY and KAHNEMAN expert and novice criterion, that is, undergraduate students are called "novices" and graduate students and PhDs in psychology are called "experts." This selection procedure was included to test the hypothesis that graduate training in statistics may not be sufficient to guarantee expertise (objective correctness) in probability calculus. The Probability

Knowledge Questionnaire was used to ascertain subjects' knowledge base of probabilistic concepts and to select groups of "novices" (subjects who score 5 or less) and "experts" (subjects who score 6 or more), in accordance with the GEBOTYS and CLAXTON-OLDFIELD criterion. All subjects completed the Probability Knowledge Questionnaire and Probability Test, consisting of ten problems each. In addition, 40 subjects received a statistical training session before answering the Probability Test problems and 40 did not receive training. There were eight groups of subjects with 10 subjects in each group.

In summary, the design of the experiment was a 2 X 2 X 2 factorial with classification according to TVERSKY and KAHNEMAN (expert versus novice), classification according to GEBOTYS and CLAXTON-OLDFIELD (expert versus novice), and training versus no training as between-subjects variables.

Method

Subjects

Subjects were 40 undergraduate students, mostly first and second year volunteers, from Wilfrid Laurier University and 40 graduate student and faculty volunteers from the Psychology Departments at Wilfrid Laurier University, the University of Waterloo, and the University of Toronto. Fifteen of these 40 subjects were PhDs and 25 were graduate students who had completed graduate level statistics course requirements at each of the schools.

Data obtained from 7 subjects were not analyzed because the classification criteria employed assigned them to groups that were filled with subjects.

Procedure

Subjects were informed that they were participating in a study of how people make judgments under uncertainty. All subjects first completed the Probability Knowledge Questionnaire. The questionnaire consisted of a 5-page booklet containing 10 problems to which various probabilistic concepts could be applied (e.g., law of large numbers, etc.). The order in which problems appeared was randomized but was the same for all subjects. Subjects were asked to select a multiple choice answer for some problems and to provide either a calculation or expression for others. Subjects were instructed to work on the problems at their own speed and not to guess at the answer to any problem, but rather to indicate "don't know" if they could not solve a problem. In addition, subjects were informed that they would receive 25 cents for each correct answer. Subjects who scored 6 or more correct out of 10 on the questionnaire were called "experts" and subjects who scored 5 or fewer were called "novices" according to the GEBOTYS and CLAXTON-OLDFIELD criterion.

According to the TVERSKY and KAHNEMAN criterion, subjects with graduate training in statistics (graduate students and faculty in psychology) were called "experts" and laypersons (undergraduate students) were called "novices." A total of 4 groups of 20 subjects each participated in the experiment. Subjects were cross-classified as experts-experts, experts-novices, novices-experts, and novices-novices.

Half (10) of the subjects in each of the four groups were randomly assigned to the training condition. The brief training session was designed to sensitize subjects to some basic probabilistic concepts. These subjects received training individually immediately after completing the Probability Knowledge Questionnaire. The training session lasted approximately 15 minutes. Subjects received a simple explanation of the proper probabilistic concepts required to solve the 10 problems that appeared on the Probability Knowledge Questionnaire plus one additional problem. The additional problem was included as an extra example for the training session. Diagrams, which could be written on for the purpose of further explaining the concepts, were used for problems 1, 2, 3, 5, 6, 7, 8, 9, and the additional problem. The training diagrams appear in Appendix D. Subjects in the no training condition were told only that they had correctly or incorrectly solved the Probability Knowledge Questionnaire problems and the additional problem. This was done on a problem by problem basis with no explanations given. The no training subjects received the Probability Test immediately after completing the Probability Knowledge Questionnaire. All subjects were informed of the subsequent Probability Test at the beginning of the experiment.

Training Condition

For each problem in turn, the proper concept was explained. For incorrectly solved problems, the subject was asked to reread the problem and to consider it for a few moments before the experimenter explained the basic principles involved. For example, for problem 1 subjects were asked to consider a coin that is tossed twice and were shown the four possible outcomes (H H, H T, T H, T T). Subjects were asked whether they would say the probability of a H is 1 if they obtained H H or the probability of a H is 0 if they obtained T T. Finally subjects were shown a diagram plotting the proportion of Hs when a coin has been tossed from 10 to 10,000 times. The probability of a H may approach .5 only after the number of repetitions of the event (tossing a coin) becomes very large; the variance in the proportion of Hs can be quite large in small samples (e.g., two toss case). The training procedure for all 10 problems and the additional problem appears in Appendix D.

Immediately following the training (no training) session, subjects were given a Probability Test consisting of 10 Tversky and Kahneman (e.g., 1974) problems that dealt with the same probabilistic concepts as on the Probability Knowledge Questionnaire (see Appendix C). Subjects were informed that they would continue to receive 25 cents for each correct answer. The test instructions were the same as for the Probability Knowledge Questionnaire. Each problem on the Probability Test was followed by the question. "Do you feel that you have applied the proper concept in this problem?" Subjects were to circle Yes or No. This additional question was included as a measure of subjects' confidence in their judgmental expertise.

Materials

Ten problems incorporating different probabilistic concepts were developed for the Probability Knowledge Questionnaire. The inverse relationship between sample size and sampling variance applied to problems 1 and 6. Randomization and the independence of sequential events applied to problem 2. Base rates and the application of Baye's Theorem applied to problems 3, 4 and 5. The concept of correlation applied to problem 7. Permutations and combinations applied to problems 8, 9 and 10.

A list of all problems appearing in Kahneman, Slovic, and Tversky's (1982) book was compiled. Ten problems dealing with similar probabilistic concepts to the problems appearing on the Probability Knowledge Questionnaire were selected verbatim from the list for the Probability Test.

Analyses

An ANOVA was carried out on the proportion of problems from the Probability Test correctly solved by each subject.

Specific Research Questions and Predictions

1) Do "experts" as compared to "novices" according to the TVERSKY and KAHNEMAN (T&K) criterion score significantly higher on the

Probability Knowledge Questionnaire? If graduate training in statistics is sufficient to guarantee expertise in probability calculus, it would be predicted that "experts" (graduate students and PhDs in psychology) would solve more problems on the Probability Knowledge Questionnaire than "novices." However, on the basis of the literature reviewed in the Training section (e.g., Tversky & Kahneman, 1971), it was predicted that there would not be a significant difference in the proportion of Probability Knowledge Questionnaire problems solved by "experts" and "novices."

2) Does the statistical training session bring about a change in how subjects think about the Probability Test problems? On the assumption that sensitizing subjects to the normative (objective) rules of probability is a reasonable approach to reducing subject's reliance on heuristics, it was predicted that the training session would serve to make a statistical approach more salient. Subjects in the training condition would correctly solve significantly more Probability Test problems than subjects in the no training condition.

3) It was predicted that the training by T&K classification interaction would not be significant. The difference in the proportion of Probability Test problems correctly solved by "experts" and "novices" (as classified by T&K) who receive training, would not be significant, that is, both groups would benefit equally from the training session. The difference

between these same two groups (as classified by T&K), who do not receive training, would not be significant on the assumption that the "experts" would rely on the same heuristics as the "novices" and be prone to the same errors (Tversky & Kahneman, 1971).

Results

The reported findings will follow the same order as in the section Specific Research Questions and Predictions with the exception of the additional analyses on the confidence ratings, individual Probability Test problems, and answers consistent with heuristics which will be mentioned at the end of the Results section.

Probability Knowledge Questionnaire Scores

A significant main effect of classification according to GEBOTYS and CLAXTON-OLDFIELD was observed on the Probability Knowledge Questionnaire scores, $F(1, 76) = 169.303$, $p < .001$. Specifically, "experts" and "novices" had mean Probability Knowledge Questionnaire scores of 6.67 and 3.85 respectively. This difference is in accordance with the selection criterion. The mean Probability Knowledge Questionnaire scores for "experts" and "novices," as classified according to the TVERSKY and KAHNEMAN criterion, showed no statistical difference ($M = 5.42$ and 5.10 respectively, $p > .05$), as predicted. The ANOVA table and a table showing the mean Probability Knowledge Questionnaire scores as a function of classification and training can be found in Appendix E.

Probability Test Scores

A significant main effect of training was observed on the Probability Test scores, $F(1, 72) = 33.758$, $p < .001$. It is clear that training increased the proportion of objectively correct answers on the Probability Test, as predicted. Specifically, subjects who received training had a mean Probability Test score of 6.22. Subjects who did not receive training had a mean Probability Test score of 4.27. The ANOVA table and a table showing the mean Probability Test scores as a function of classification and training can be found in Appendix F.

A significant classification according to TVERSKY and KAHNEMAN by training interaction was observed on the Probability Test scores, $F(1, 72) = 4.350$, $p < .05$. Table 1 shows the mean Probability Test scores for the TVERSKY and KAHNEMAN "experts" and "novices," as a function of training.

Table 1

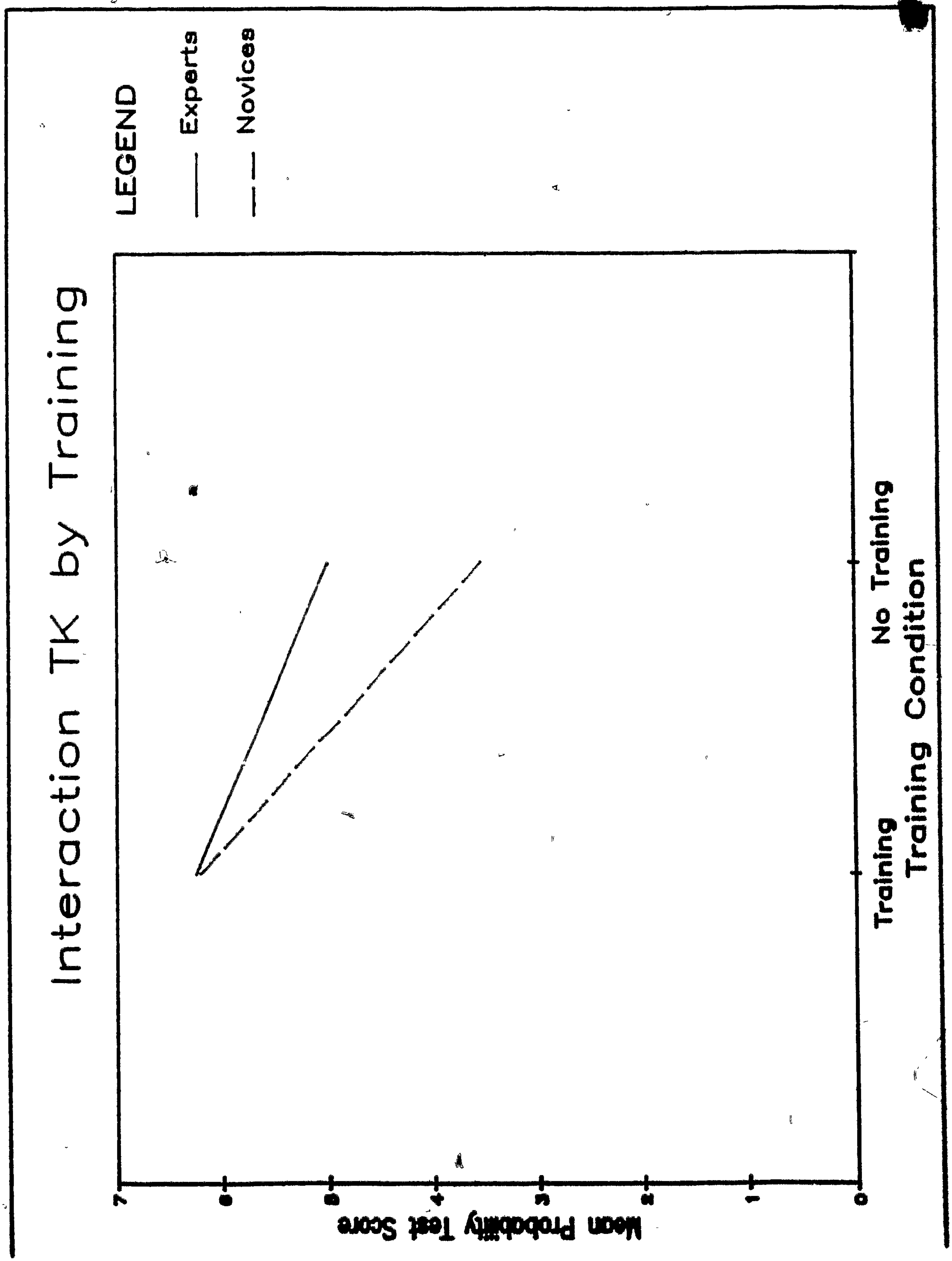
<u>Mean Probability Test Scores as a Function of Classification According to TVERSKY and KAHNEMAN and Training</u>

Condition	Experts	Novices
Training	6.25	6.20
No Training	5.00	3.55

<u>Note</u>. Maximum score = 10

It is clear that training increased the salience of a statistical approach for both groups of subjects ("experts" and "novices"). The mean Probability Test scores for the "experts" and "novices" who received training were 6.25 and 6.20, respectively. This difference was not significant, as predicted. The mean Probability Test scores for the "experts" and "novices" who did not receive training were 5.00 and 3.55 respectively. These values are significantly different and significantly different from the corresponding values in the training condition (mean comparisons via Fisher's Least Significant Difference, <u>LSD</u> = .949, $p < .05$). It is worth noting that the mean Probability Test score for the "novices" who received training (M = 6.20) was significantly higher than the mean Probability Test score for the "experts" who did not receive training (M = 5.00). Finally, the "experts" who did not receive training correctly solved a significantly higher proportion of Probability Test problems than the "novices" who did not receive training (M = 5.00 and 3.55 respectively). This interaction is presented graphically in Figure 1.

Figure 1


Interaction TK by Training
LEGEND
Experts
Novices
Mean Probability Test Score
0
1
2
3
4
5
6
7
Training
No Training
Training Condition

The Probability Test confidence proportions were compared across separate contingency tables for the training and no training conditions by crossing correct and incorrect answers with subjects' indications of how confident they were that they had applied the proper concept on each problem. A measure of subjects' confidence in their judgments was obtained by dividing the number of correct answers to problems on which subjects indicated confidence by the total number of correct answers. For example, in the training condition, subjects indicated that they were confident on 96.8% of their correct answers. In the no training condition, subjects indicated that they were confident on 92.9% of their correct answers. A comparison of the binomial proportions for confidence in correct answers showed that the proportion of correct answers on which subjects had expressed confidence was significantly higher in the training condition than in the no training condition, $z(\text{inf}) = 1.83$, $p < .05$ (using a one-tailed test).

A further measure of subjects' confidence in their judgments was obtained by dividing the number of incorrect answers to problems on which subjects indicated confidence by the total number of incorrect answers. In the training condition, subjects indicated that they were confident on 81.1% of their incorrect answers. Subjects in the no training condition indicated that they were confident on 86.1% of their incorrect answers. A comparison of the binomial proportions for confidence in incorrect answers showed no significant difference between the proportion of incorrect answers on which subjects had expressed confidence in the training and no training conditions, $z(\inf) = -1.17$, $p < .05$ (using a one-tailed test).

A measure of subjects' "accuracy" in their judgments was obtained by summing the number of correct answers to problems on which subjects indicated that they were confident and the number of incorrect answers to problems on which subjects indicated that they were not confident. Accuracy was defined as the level of agreement between subjects' answers and their indicated confidence. For example, in the training condition, subjects were correct and confident for 240 problems and incorrect and not confident for 24 problems. Thus, subjects' accuracy of judgments matched their confidence in judgments for 265 (70.7%) of the problems. In the no training condition, subjects' accuracy of judgments matched their confidence in judgments for 182 (53.1%) of the problems. Overall, confidence ratings were missing or subjects indicated don't know for 82 of the 800 problems (10.3%). A comparison of the binomial proportions for accuracy (70.7% versus 53.1%) showed that a significantly higher proportion of judgments were made more accurately in the training condition than in the no training condition, $z(\inf) = 4.76$, $p < .001$. The data are summarized in Table 2.

Table 2

Probability Test Confidence Proportions as a Function of Training

		Confidence	
Condition	Answer	Yes	No
Training	Correct	240	8
		64.00%	2.13%
	Incorrect	103	24
		27.46%	6.40%
No Training	Correct	158	12
		46.10%	3.50%
	Incorrect	149	24
		43.40%	6.90%

The results of separate analyses of variance on the Probability Test problems are summarized in Appendix G. A comparison of the proportion of subjects who correctly answered each problem, as a function of training, yielded significant differences on problems 2, 3, 4, 7, 8, 9, and 10. Although not significant, the differences on problems 1 and 6 were in the expected direction.

Finally, the proportion of answers consistent with the representativeness and availability heuristics were compared for the training and no training conditions. A measure of subjects' reliance on heuristics was obtained by dividing the number of answers to problems that were consistent with heuristics by the total number of answers. In the training condition, subjects' answers were consistent with heuristics for 76 of the problems. Thus, subjects in the training condition relied on heuristics for 19% of the problems. In the no training condition, subjects' answers were consistent with heuristics for 114 (29%) of the problems. A comparison of the binomial proportions for reliance on heuristics showed that a significantly higher proportion of answers were consistent with the representativeness and availability heuristics in the no training condition than in the training condition, $z(inf) = -3.33$, $p < .001$ (using a one-tailed test).

It should be noted that it was not practical to compare the present findings, on a problem by problem basis, with the findings reported by Kahneman and Tversky (e.g., 1972). Kahneman and Tversky primarily reported median estimates and modal responses for their problems, whereas mean responses were reported in the present study. Subjects in the Kahneman and Tversky studies were primarily high-school students, whereas those in the present study were recruited from a university.

Discussion

The findings of the experiment designed to investigate the effects of training on the quantification of uncertainty may be summarized as follows. It is argued that the 15-minute training session served to make a statistical approach to the Probability Test problems more salient. This is not to imply that the training session turned subjects into instant statisticians, but rather gave them a minimal probability knowledge base. The brief training session enabled subjects to correctly apply probability calculus to a set of "difficult problems" (Fong et al., 1986, p. 281). These were problems to which previously no subjects, or almost no subjects applied statistical principles. The training not only served to sensitize subjects to the probabilistic concepts, but may have enhanced subjects' accuracy in assessing their answers to the Probability Test problems.

It was observed that the 15-minute training session served to increase the proportion of objectively correct answers to the Probability Test problems. It has been reported that subjects' judgments about the birth-sequence problem, engineer-lawyer problem and so on frequently deviate from objectively correct answers (e.g., Tversky & Kahneman, 1974). According to Tversky and Kahneman (1974), this is due to the application of heuristics (i.e., representativeness and availability). This phenomenon is seen as robust since experts (e.g., professional psychologists) make use of the same heuristics as laypeople (Tversky & Kahneman, 1971). However, subjects who received the 15-minute training session were more likely to apply the normative (objectively correct) probabilistic concepts to the Tversky and Kahneman problems on the present study's Probability Test. These preliminary findings suggest that what is claimed to be a robust phenomenon (i.e., reliance on heuristics) may in fact be a weak one.

The Fong et al. (1986) training studies demonstrated that statistical training served to enhance the use of statistical principles in reasoning about everyday problems. Subjects' answers to these problems were coded for the presence and quality of a statistical response. That is, subjects were not presented with problems that had a "correct" answer. The present investigation studied the effects of statistical training from a normative standpoint, emphasizing objectively correct judgments. Decomposing the task to simple units of understanding and making explicit the concepts necessary to solve the problems appeared to be a reasonable approach to reducing subjects' reliance on the representativeness and availability heuristics.

Various interpretations may be offered to account for the pattern of findings reported by Kahneman and Tversky (1972, 1973, 1982; Tversky & Kahneman, 1971, 1974).

a) Subjects used the representativeness and availability heuristics to make their judgments although they understood the normative probabilistic concepts. This would be evidence for the robustness of heuristics.

b) Subjects used heuristics that deviated from both objectively correct solutions and the hypothesized representativeness and availability heuristics (Olson, 1976).

c) Subjects may not understand the normative probabilistic concepts. This interpretation suggests that if you sensitize people to the proper application of the concepts you can reduce their reliance on heuristics.

The latter is the most compelling for a number of reasons. Firstly, it may be that Kahneman and Tversky's subjects were not sensitized to the normative concepts of probability. Most of Kahneman and Tversky's subjects, with the exception of the professional psychologists (1971), were high-school students. It is not surprising that these statistically unsophisticated subjects did poorly and gave rather atypical answers in Kahneman and Tversky's studies. Secondly, the present study's preliminary findings indicate that graduate training in psychological statistics is not sufficient to guarantee expertise in probability calculus or statistics. The results of an ANOVA comparing the mean Probability Knowledge Questionnaire scores of "experts" and "novices", as classified on the TVERSKY and KAHNEMAN factor, showed no statistical differences. Thirdly, the TVERSKY and KAHNEMAN "novices," who received training, correctly solved a significantly higher proportion of Probability Test problems than the TVERSKY and KAHNEMAN "experts" without training. This result suggests that the TVERSKY and KAHNEMAN "experts" who did not receive training may not have had an adequate probability knowledge base.

Training served to enhance subjects accuracy in assessing whether or not they had applied the proper concepts to the Probability Test problems. However, this finding may be due to differences in the number of correct answers in the training and no training conditions. A more appropriate measure of accuracy would have been to have subjects indicate on each problem whether they believed their answer was correct or not before they indicated whether or not they were confident that they had applied the proper concept to the problem. Training served to enhance subjects' confidence that they had applied the proper concepts to the Probability Test problems on which they were correct. Contrary to the previous findings (e.g., Oskamp, 1965), subjects who received training did not become overconfident in their judgments relative to subjects who did not receive training. That is, although subjects who received training indicated more confidence in their correct answers, their confidence ratings in their incorrect answers were not out of proportion to the ratings of subjects who did did not receive training. However, consistent with the previous literature (e.g., Hoch, 1985), subjects were overconfident in their true judgmental abilities, as evidenced by the proportion of incorrect answers on which subjects indicated they were confident.

The preliminary study reported here demonstrated that training served to make a statistical approach to the Probability Test problems more salient. It is important to note that the training effects in the present study did not appear to be in the form of a memory about how to "map" the concepts onto the Probability Test problems. That is, it did not appear to be the case that subjects simply memorized the answers to the Probability Knowledge Questionnaire problems and strictly applied the same answers to the Probability Test problems. Had this been the case, it could be argued that the training did not teach subjects anything new at all. To test this "memory" hypothesis, two independent reviewers rated the similarity of the Probability Knowledge Questionnaire problems to the Probability Test problems. Both reviewers indicated that problem 1 on the Probability Knowledge Questionnaire was similar to problems 1 and 6 on the Probability Test involving the the concept of sample size and sampling variance and that problems 3, 4, and 5 on the Probability Knowledge Questionnaire were dissimilar to problems 4 and 10 on the Probability Test involving the concept of base rates. A "memory" hypothesis would predict a significant effect of training on problems 1 and 6 on the Probability Test due to the similarity of the training problem to the test problems. That is, subjects memorized the law of large numbers and applied the principles to problems 1 and 6 on the Probability Test because these problems were similar to the problem on which the law of large numbers was trained (e.g., subjects simply responded to stated sample sizes). A "memory" hypothesis would not predict a significant effect of effect of training on problems 4 and 10 due to the dissimilarity of the training problems to the test problems. For example, the concept trained on the ball and urn problems was tested in substantially broader domains, that is, the domain of judgments about the description of a person and the cab problem. This fact could be expected to reduce "memory" effects to a minimum. However, the results of separate ANOVAs on the Probability Test problems showed no significant training effects on problems 1 and 6, but a significant effect of training was observed on problems 4 and 10. These results are inconsistent with a strict "memory" hypothesis, but are consistent with a minimal probability knowledge base hypothesis. Due to the temporal relationship between testing and training, it could be expected that subjects would have the concepts in "active memory" at the time they were asked to answer the Probability Test problems. However, the present findings suggest that subjects understood the applicability of some concepts even when the Probability Test problem which tested a particular concept was dissimilar to the Probability Knowledge Questionnaire problem on which the concept had been trained.

It is interesting to note the apparent limitations of heuristics. The preliminary findings reported here demonstrated that reliance on heuristics can be reduced as a result of a 15-minute training session. It appears that people use heuristics to make judgments when they do not know how to apply the normative probabilistic concepts to the problems. This is not to imply that the present training strategy was successful in producing correct answers to all problems. There were some inconsistencies in the data. For example, a significant main effect of training was observed for the medical survey problem (problem 3), but not for the maternity-ward and pollster problems (problems 1 and 6), although all three problems involved the application of the law of large numbers. For example, the maternity-ward problem was a difficult problem for subjects. It can be argued that this problem was less transparent (i.e., more complex) than the pollster problem. The correct answer to the maternity-ward problem can be elicited by recognizing that a day on which more than 60% of babies born are boys is a departure from the ideal that about 50% of all babies are boys. And so, if one has greater confidence in a larger sample, one should expect that the smaller hospital will record more days on which over 60% of the babies born are boys. Overall, however, the training condition did serve to increase the proportion of correct answers to the Probability Test problems relative to the no training condition. Future studies could better explore individual problems by including larger samples of subjects or a smaller subset of problems dealing with only one concept (e.g., Fong et al., 1986). It was also observed that the difference in the proportion of problems correctly solved by TVERSKY and KAHNEMAN's "experts" and "novices" was not significant on the Probability Knowledge Questionnaire, but was significant on the Probability Test. One explanation for this inconsistency is that some of the "experts" (PhDs and graduate students) had been exposed to some of the Kahneman and Tversky problems in their course training. This was mentioned to the experimenter by a few subjects. Other limitations of the present preliminary study should also be noted. First, future research should include larger and random samples of PhDs and graduate students to ensure that more representative samples be obtained. Secondly, the present study had two major conditions: a control group given no training and an experimental group trained in probability calculus. Although it is clear that the training had a significant effect on Probability Test performance and provided preliminary information on why people use heuristics and their robustness, it is not clear what specific features of the training session were successful in sensitizing subjects to the concepts tested. It may be that the diagrams were more effective than the verbal explanations or that the combination of both was more effective than either alone. Future research could assess the effectiveness of different kinds of training to see if there is an effect. Finally, the training effects occurred when the concepts were tested immediately after training. Although the preliminary findings indicate that reliance on heuristics can be reduced with training, future research could assess the effects of training on the use of heuristics when the temporal relation between training and testing is delayed in time.

References

Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. Journal of Personality and Social Psychology, 33, 303-314.

Bar-Hillel, M. (1982). Studies of representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs. concrete information on decisions. Journal of Applied Social Psychology, 7, 258-271.

Carroll, J. S., & Siegler, R. S. (1977). Strategies for the use of base-rate information. Organizational Behavior and Human Performance, 19, 392-402.

Cohen, C. E. (1981). Person categories and social perception: Testing some boundaries of the processing effects of prior knowledge. Journal of Personality and Social Psychology, 40, 441-452.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. Psychological Review, 82, 407-428.

Darley, J. M., & Latane, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. Journal of Personality and Social Psychology, 8, 377-383.

DuBois, N. F., Alverson, G. F., & Staley, R. K. (1979) Educational psychology and instructional decisions. New York: The Dorsey Press.

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

Fischhoff, B., & Beyth, R. (1975). "I knew it would happen" - Remembered probabilities of once-future things. Organizational Behavior and Human Performance, 13, 1-16.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986) The effects of statistical training on thinking about everyday problems Cognitive Psychology, 18, 253-292.

Fraser, D. A. S. (1967). Statistics: An introduction. New York: John Wiley & Sons, Inc.

Gebotys, R. J., & Claxton-Oldfield, S. P. (1986). Probability knowledge questionnaire. Unpublished questionnaire. Wilfrid Laurier University, Waterloo.

Ginosar, Z., & Trope, Y. (1980). The effects of base rates and individuating information on judgements about another person. Journal of Experimental Social Psychology, 16, 228-242.

Hamill, R., Wilson, T. D., & Nisbett, R. E. (1980). Insensitivity to sample bias: Generalizing from atypical cases. Journal of Personality and Social Psychology, 39, 578-589.

Hoch, S. J. (1985). Counterfactual reasoning and accuracy in predicting personal events. Journal of Experimental Psychology: Learning, Memory and Cognition, 11, 719-731.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgement of representativeness. Cognitive Psychology, 3, 430-454.

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

Kahneman, D., & Tversky, A. (1979). Intuitive prediction: Biases and corrective procedures. TIMS Studies in Management Science, 12, 313-327.

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

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. Journal of Experimental Psychology: Human Learning and Memory, 4, 551-578.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97

Manis, M., Dovalina, I., Avis, N. E., & Cardoze, S. (1980). Base rates can affect individual predictions. Journal of Personality and Social Psychology, 38, 231-248.

Nahinsky, I. D., & Ash, D. (1985). Unidirectional independence and judgmental heuristics. Bulletin of the Psychonomic Society, 23, 467-469.

Nisbett, R. E., & Schacter, S. (1966). Cognitive manipulation of pain. Journal of Experimental Social Psychology, 2, 227-236.

Nisbett, R. E., & Borgida, E. (1975). Attribution and the psychology of prediction. Journal of Personality and Social Psychology, 32, 932-943.

Nisbett, R. E., Borgida, E., Crandall, R., & Reed, H. (1976). Popular induction: Information is not necessarily informative. In J. Carroll & J. Payne (Eds.), Cognitive and social behavior. Hillsdale, N.J.: Erlbaum.

Nisbett, R. E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgement. Englewood Cliffs, N.J.: Prentice-Hall.

Nisbett, R. E., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the effect of diagnostic information. Cognitive Psychology, 13, 248-277.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Fong, G. T. (1982). Improving inductive inference. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press.

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.

Olson, C. L. (1976). Some apparent violations of the representativeness heuristic in human judgment. Journal of Experimental Psychology: Human Perception and Performance, 12, 599-608.

Oskamp, S. (1965). Overconfidence in case-study judgments. The Journal of Consulting Psychology, 29, 261-265.

Robinson, G. H. (1964). Continuous estimation of a time-varying probability. Ergonomics, 7, 7-21.

Rosch, E. (1975). Cognitive representation of semantic categories. Journal of Experimental Psychology: General, 104, 192-253.

Rosch, E., & Mervis, C. B. (1975). Family resemblance: Studies in the internal structure of categories. Cognitive Psychology, 7, 573-605.

Shuford, E. H. (1961). Percentage estimation of proportion as a function of element type, exposure time, and task. Journal of Experimental Psychology, 61, 430-436.

Slovic, P., & Combs, B. (1979). Newspaper coverage of causes of death. Journalism Quarterly, Winter.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 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. (1974). Judgement under uncertainty. Science, 185, 1124-1131.

Weber, J. D. (1973). Historical Aspects of the Bayesian Controversy. Division of Economics and Business Research, College of Business and Public Administration, University of Arizona, Tucson.

Zukier, H. (1982). The dilution effect: The role of correlation and the dispersion of predictor variables in the use of nondiagnostic information. Journal of Personality and Social Psychology, [illegible], 1163-1174.

Appendix A

Pilot Study Questionnaire

2 of/de 2

The following description of an individual was randomly drawn from a sample of 100 descriptions of which 70 were of lawyers and 30 were of engineers:

Dan is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

The probability that Dan is one of the 70 lawyers in the sample of 100 is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very Not Very

A coloured ball was randomly drawn from an urn containing a sample of 100 balls of which 70 were blue and 30 were red.

The probability of drawing one of the blue balls from the sample of 100 is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very   Not Very

The following description of an individual was randomly drawn from a sample of 100 descriptions of which 70 were of lawyers and 30 were of engineers:

Dan is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleages.

The probability that Dan is one of the 30 engineers in the sample of 100 is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very   Not Very

A coloured ball was randomly drawn from an urn containing a sample of 100 balls of which 30 were blue, 48 were red, and 22 were white.

The probability of drawing one of the red balls from the sample of 130 is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very   Not Very

Determine the probability that at least two people in a group of 23 randomly selected people share the same birthday (i.e., same month and day). Assume that there are 365 days in a year (i.e., not a leap year).

The probability that at least two people in a group of 23 people share the same birthday is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very   Not Very

Six married couples are standing in a room. If 4 people are selected at random, determine the probability that 2 married couples are selected.

The probability that 2 married couples are selected is _____%.

How confident are you of your probability estimate (circle one):

--- 1 --- : --- 2 --- : --- 3 --- : --- 4 --- : --- 5 ---

Very   Not Very

Describe in your own words what is meant by each of the following statements:

The probability of precipitation tomorrow is .3

The probability of a particular student graduating from W.L.U. is .9

The probability of a particular horse winning the third race at Woodbine is .06

The probability of winning the lottery is .0000008

The probability of an odd number appearing in a single toss of a fair die is .5

The probability of a major storm touching down in exactly the same place twice is 0

The probability of being the victim of a violent crime such as murder, rape, robbery, or aggravated assault within the next six months is .02

Appendix B

Probability Knowledge Questionnaire

1. Consider a coin, symmetric and delicately labeled Head on one side and Tail on the other. Toss the coin 10 times and record the number of Heads. Toss the coin 100 times and record the number of Heads. In which case would you expect the proportion of Heads to be closer to .5?

   a) 10 toss case
   b) 100 toss case
   c) both equally likely
   d) don't know

2. Toss a symmetric, delicately labeled coin (labeled Head on one side and Tail on the other) 5 times. Two possible sequences are given below.

   1) H H H H H

   2) H T T H T

   Is

   a) 1 most plausible
   b) 2 most plausible
   c) both equally plausible
   d) don't know

3. Consider 2 urns containing red and black balls, identical except for colour.

Urn I	Urn II
80 red	10 red
20 black	90 black

Select a ball at random from urn I.

What is the probability of obtaining a red ball?

4. Say an urn was chosen at random and a red ball is drawn. What is the probability that the ball was chosen from urn I?

5. Say that urn I weighs 100 lbs and urn II weighs 20 lbs. An urn is chosen at random by a technician. However, the technician must carry the urn 50 ft before a ball is drawn. If the ball chosen is red, calculate the probability the ball is chosen from urn I.

Would you answer question 5 differently if you knew that the technician was an athlete?

6. On the average, how many times must a die be thrown until one gets a 6?

7. X and Y are positively correlated. Would you have more confidence in predicting

a) X from Y
b) Y from X
c) equal confidence in a or b
d) don't know

8. Consider the word DOGS. How many sequences, with repetition of letters, can one form with these 4 letters?

9. In the word DOGS, how many sequences, without repetition of letters, can one form?

10. A 3 man jury has 2 members each of whom independently has a probability p of making a correct decision and a third member who flips a coin for each decision (majority rules).

A 1 man jury has probability p of making the correct decision.

Which jury has the higher probability of making a correct decision?

a) the 3 man jury
b) the 1 man jury
c) both equally
d) don't know

Additional Problem

Consider choosing people and record the amount of money they carry. It is known that the number of people who carry more than $20.00 is equal to the number of people who carry less than $20.00.

Say you.

1) pick one person and record the amount of money he or she carries.

2) pick three people and determine the amount of money they carry. Order the amounts from smallest to largest and record the middle amount.

If you were to repeat the above procedures many times, would

a) Method 1 have more people with cash greater than $30.00
b) Method 2 have more people with cash greater than $30.00
c) both equal
d) don't know

Appendix C

Probability Test

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

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

a) the larger hospital
b) the smaller hospital
c) about the same
d) don't know

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

2. All families of six children in a city were surveyed. In 72 families the exact order of births of boys (B) and girls (G) was G B G B B G.

What is your estimate of the number of families surveyed in which the exact order of births was B G B B B B?

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

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

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

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

4. A panel of psychologists has interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find below a single description chosen at random from the 100 available descriptions. Please indicate your probability that the person described is a lawyer.

Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

The probability that Jack is one of the 70 lawyers in the sample of 100 is:

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

5. The average height of American college men is 175 cm. Three files were randomly drawn from a college registrar's office, belonging to John, Mike, and Bob. Which outcome is more likely with respect to the heights of these three men?

a) John (178 cm), Mike (170 cm), Bob (176 cm)
b) John (177 cm), Mike (177 cm), Bob (177 cm)
c) both equally likely
d) don't know

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

6. Two pollsters are conducting a survey to estimate the proportion of voters who intend to vote YES on a certain referendum. Firm A is surveying a sample of 400 individuals. Firm B is surveying a sample of 1,000 individuals. Whose estimate would you be more confident in accepting?

a) Firm A's
b) Firm B's
c) about the same
d) don't know

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

7. In which prediction would you have greater confidence?

a) the prediction of a man's weight from his height
b) the prediction of a man's height from his weight
c) equal confidence in a or b
d) don't know

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

8. Consider the following two structures. A and B, which are displayed below.

(A)
X X X X X X X X
X X X X X X X X
X X X X X X X X

(B)
X X
X X
X X
X X
X X
X X
X X
X X
X X

A path in a structure is a line that connects an element in the top row to an element in the bottom row, and passes through one and only one element in each row.

In which of the structures are there more paths?

a) Structure A
b) Structure B
c) the same in A and B
d) don't know

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

9. How many paths do you think there are in each structure?

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

10. 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: (a) 85% of the cabs in the city are Green and 15% are Blue. (b) a witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each of the two colours 80% of the time and failed 20% of the time.

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

Do you feel that you have applied the proper concept in this problem? (Circle one)

Yes No

Appendix D

Training Procedure and Diagrams

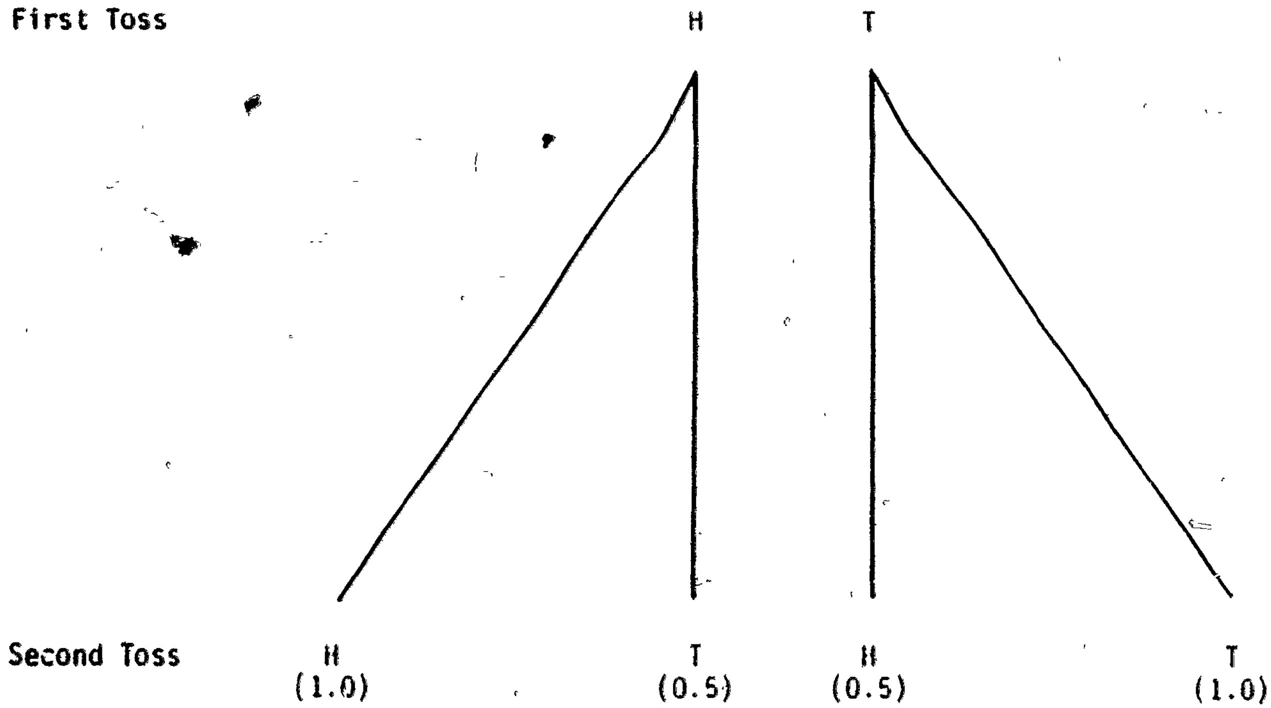
Problem 1

If you toss a coin in the air, you are certain it will come down, but you are not certain that, say, a Head will appear. We use probability to quantify uncertainty. Problem 1 asks in which case would you expect the proportion of Heads to be closer to .5. Most people know that when the probability of an event occurring is .5 then half of the time, on average, the event will occur. However, fewer people realize that the average may approach 50 percent only after a large number of independent repetitions of the event.

Consider the event tossing a coin twice. [Subjects are shown diagram (Diagram 1) of two toss case] The event of interest is the proportion of Heads. On the first toss, you can have a H or T and on the second toss, you can have a H or T The 4 possible outcomes are ... [write on diagram H H, H T, T H, and T T] If you toss H on the first toss and H on the second toss, the proportion of Heads is 1; if you toss T on the first toss and T on the second toss, the proportion of Heads is 0. However, if you toss H on the first toss and T on the second toss or T on the first toss and H on the second toss, the proportion of Heads is .5. In the short run, that is, with a two toss case, the variance in the proportion of Heads is quite large.

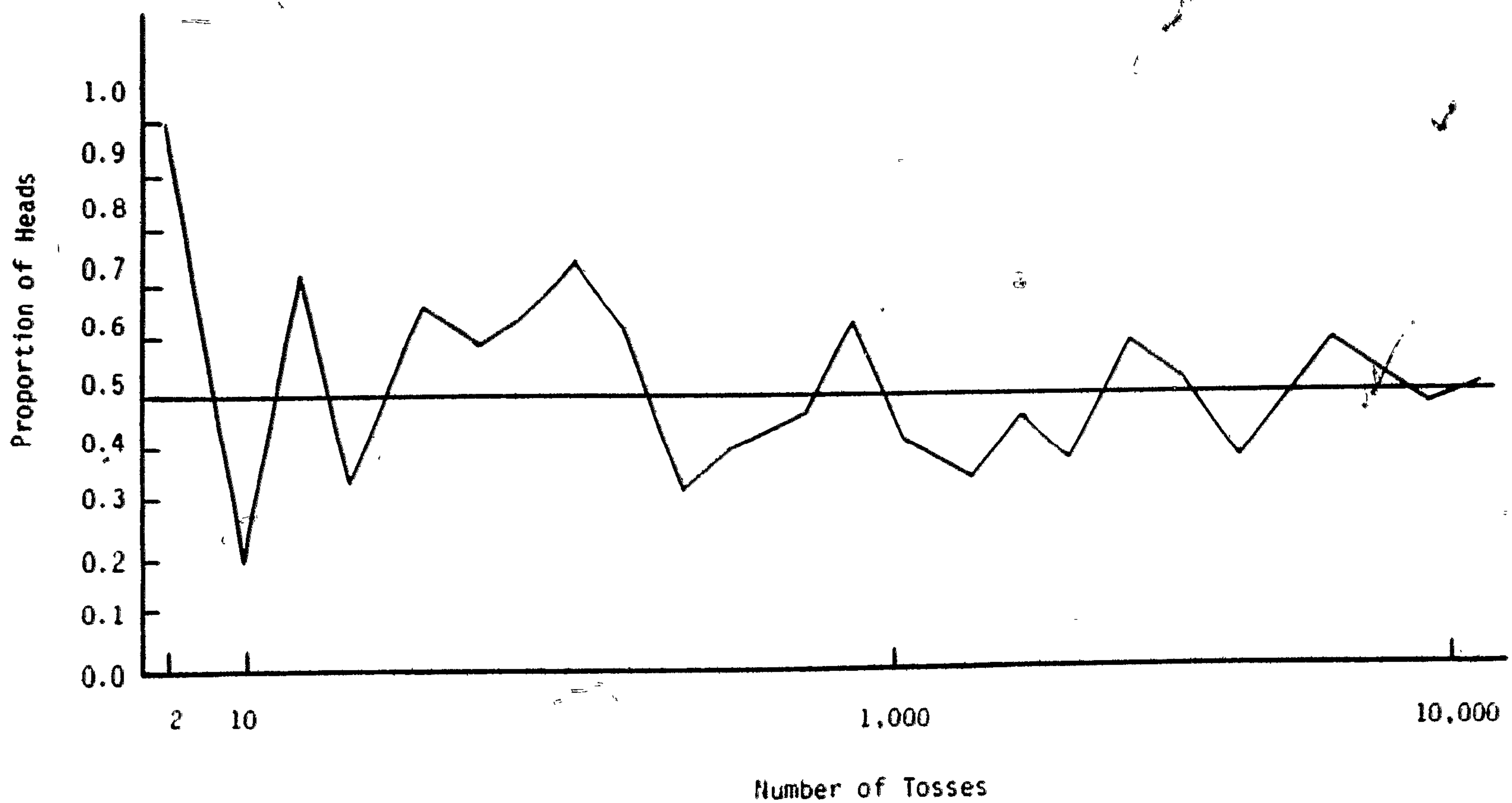
Consider the diagram plotting the proportion of Heads as a function of the number of times a coin has been tossed (Diagram 2). The proportion of Heads is determined by dividing the number of Heads observed by the total number of tosses. It has been empirically observed that the ratio the number of Heads/the total number of tosses becomes stable in the long run. That is to say, with a very large number of tosses, ~~the ratio~~ approaches a constant. The constant value is the probability assigned to the event tossing a Head. However, in the short run, that is, with a very small number of tosses, large variations in the proportion of Heads can occur. For example, with the 2 toss case the proportion of Heads was 1 or 2 Heads out of 2 tosses; with the 10 toss case the proportion of Heads was .2 or 2 Heads out of 10 tosses. With the 1000 toss case, the proportion of Heads was .4 or 400 Heads out of 1000 tosses and with the 10,000 toss case the proportion of Heads was .49 or 4,900 Heads out of 10,000 tosses. As you can see, the proportion of Heads clusters around the constant $p = .5$ as the number of tosses in the sample increases.

Proportion of Heads (Hs) with a two-toss case



Four possible outcomes:

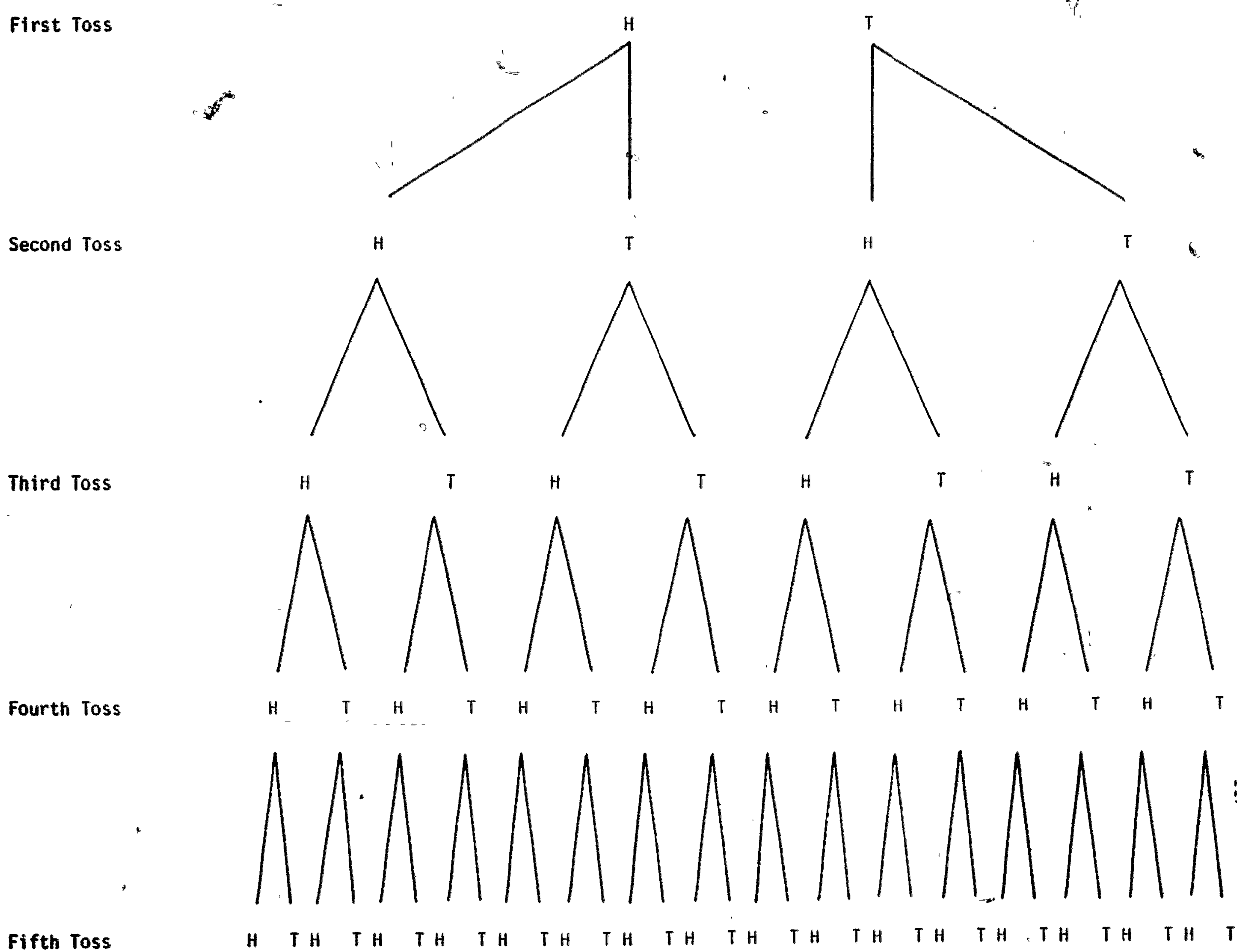
The record of 10,000 tosses of an ideal coin (proportion of Heads)



Problem 2

The events of interest in problem 2 are the number of possible sequences of Heads and Tails not the number of Heads or Tails in any particular sequence. There are two possible outcomes on any coin toss (Head or Tail). The outcomes of coin tosses are independent of each other. That is, whether a Head or Tail appeared on the first toss will not influence whether a Head or Tail will appear on the second toss and so on. [Show subjects diagram (Diagram 3) of 5 toss case]

Consider the diagram. The first toss can be Head or Tail. If the first toss is Head, the second toss can be Head or Tail and so on. If you toss a coin 5 times, there are 2 X 2 X 2 X 2 X 2 = 32 possible outcomes. [The 32 possible sequences are shown on the diagram]. If you were to repeatedly toss a coin 5 times and observe the sequences of Heads and/or Tails that appear, then each sequence would appear, on average, with equal probability or 1 in 32 times.


First Toss
H T
Second Toss
H T H T
Third Toss
H T H T H T H T
Fourth Toss
H T H T H T H T H T H T H T H T
Fifth Toss
H T H T H T H T H T H T H T H T H T H T H T H T H T H T H T H T

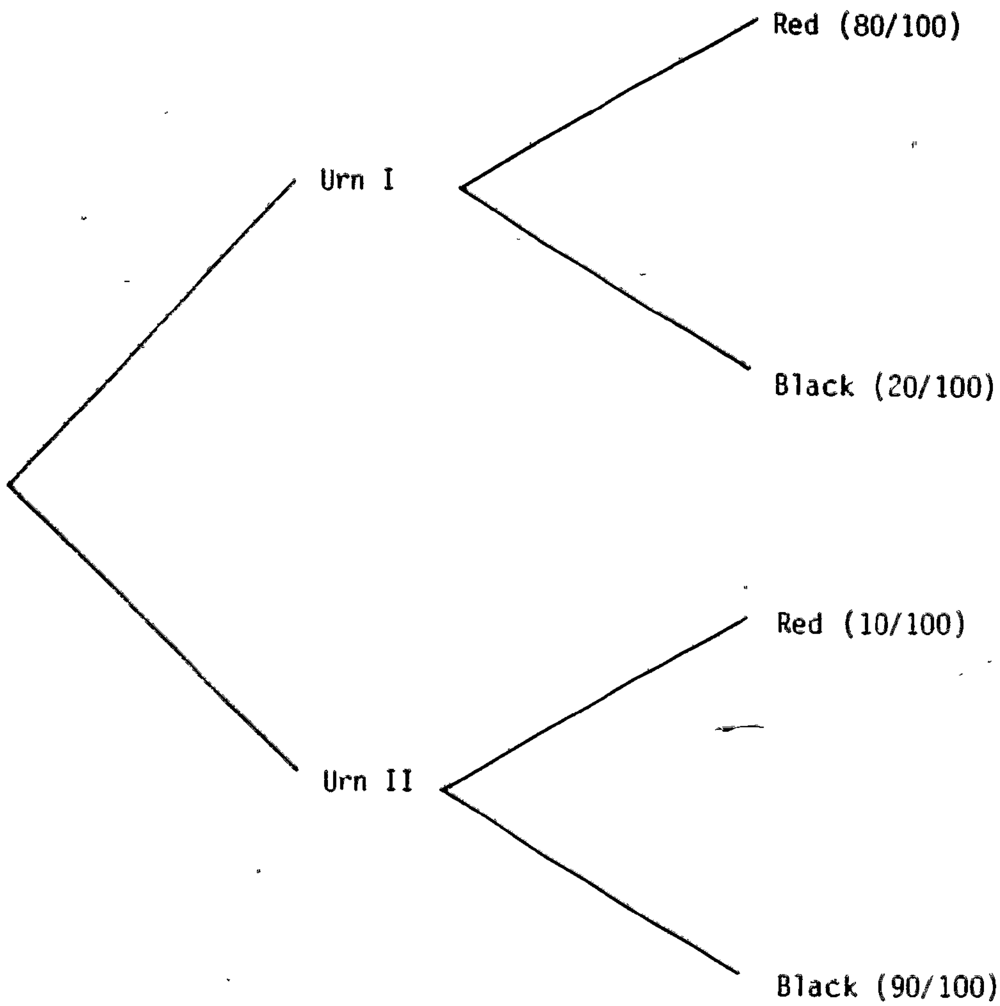
Problem 3

Urn I contains a total of 100 coloured balls. Of these 100 balls, 80 are red and 20 are black. The total possibilities are that the ball drawn from Urn I is red or black. Because the total possibilities are red or black, we add the number of red and black balls. Therefore, the total possibilities are 80 red balls + 20 black balls = 100. The event of interest in problem 3 is drawing a red ball from Urn I. There are 80 red balls in Urn I. On average then, you would expect to draw a red ball from Urn I 80/(80 + 20) or 80/100 times which equals .8.

Problem 4

The total possibilities are that the ball is drawn from Urn I or Urn II. The Urn was chosen at random. If the process whereby the Urn was selected is truly random, then half of the time, on average, Urn I will be chosen and half of the time Urn II will be chosen. To randomly choose an Urn you might consider tossing a coin; if Head then choose Urn I and if Tail then choose Urn II. In the long run, you would expect to choose Urn I half of the time and Urn II half of the time. [Show subjects diagram (Diagram 4)]

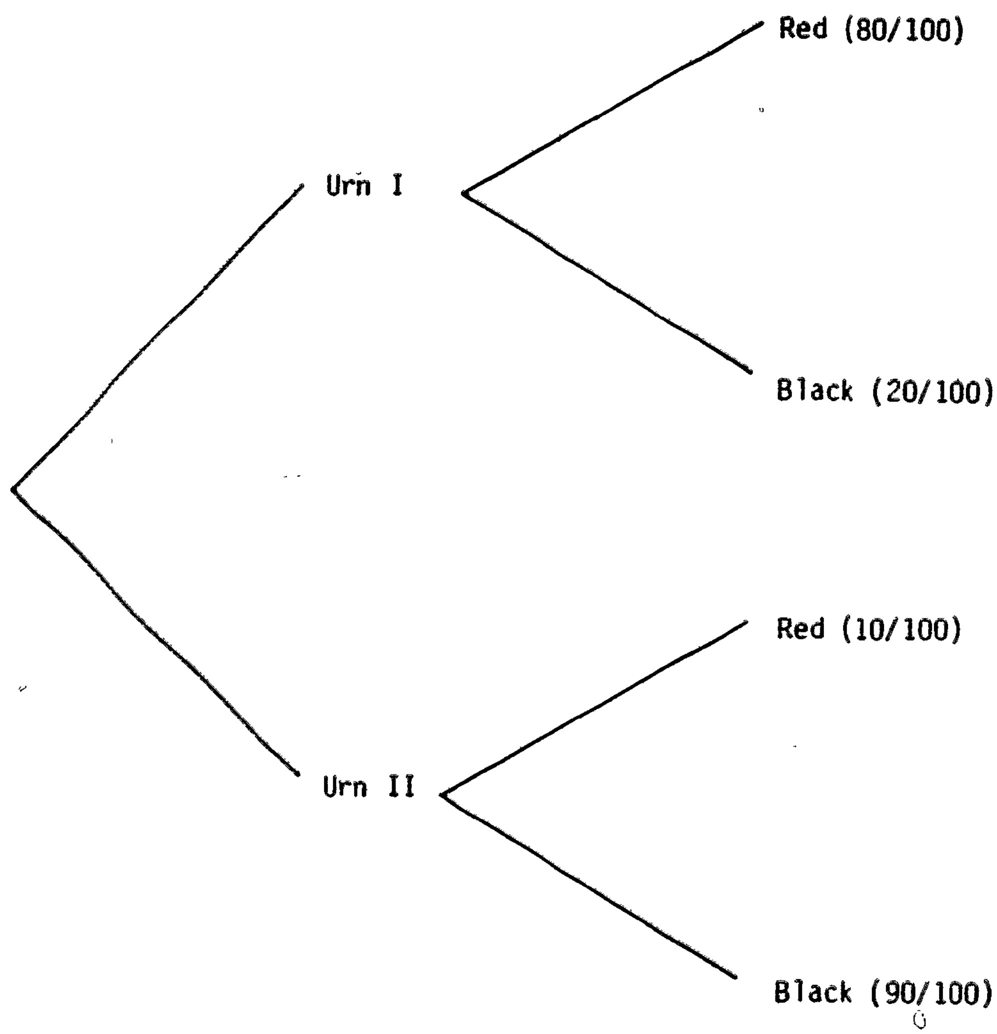
The event of interest in problem 4 is a red ball chosen from Urn I. The total possibilities are Urn I or Urn II. All you know is that a red ball was drawn, you don't know from which Urn. The sample space then is Urn I and red or Urn II and red. The probability of selecting a red ball from Urn I equals 80/(100). The probability of selecting a red ball from Urn II equals 10/(100). A correction factor is included to account for randomness We know that, on average, Urn I will be chosen 1/2 of the time and Urn II will be chosen 1/2 of the time. We multiply the probability of choosing a red ball from Urn I by 1/2 and we multiply the probability of choosing a red ball from Urn II by 1/2. Because the red ball could be chosen from Urn I or Urn II, we add the two to obtain the total possibilities of choosing a red ball. The event of interest is choosing a red ball from Urn I. You know from problem 3 that the probability of drawing a red ball from Urn I equals 80/(80 + 20). You multiply the probability by 1/2 because of randomness. The solution is obtained by dividing the event of interest by the total possibilities.


Red (80/100)
Urn I
Black (20/100)
Red (10/100)
Urn II
Black (90/100)

Problem 5

Problem 5 is similar to problem 4 except you are given additional information about the Urns, that is, Urn I weighs 100 pounds and Urn II weighs 20 pounds and the technician must carry the Urn 50 feet. If you feel that there should be a change in the correction factor to include other information then change the correction factor of 1/2 (assigned because of randomness) in problem 5. [Show subjects diagram (Diagram 5)]. However, changing the correction factor is not an arbitrary matter. The correction factor may be changed to include other information but you cannot ignore the basic information, that is, the number of red and black balls in the Urns. For example, if you feel the technician might tire carrying a 100 pound Urn 50 feet repeatedly before drawing a ball, you might want to assign a smaller correction factor to Urn I than to Urn II, say 1/4. Although selection of an Urn is random, you may feel that the technician might go to the 20 pound Urn more because he has a sore back and doesn't want to carry the 100 pound Urn too many times. Assigning a lower correction factor to the heavier Urn is subjective, i.e., you may feel the technician isn't honest. If Urn I is assigned a correction factor of 1/4 indicating that you think the technician will go to Urn I, on average 1 in 4 times rather than 1 in 2 times, then Urn II is assigned a correction factor of (1 - 1/4) = 3/4. Here are some calculations using the basic information but using different correction factors for each Urn (Diagram 6).

You don't have to change the correction factor at all. You may use 1/2 as in problem 4. There are no right or wrong answers for problem 5 provided the basic information is not ignored.


Red (80/100)
Urn I
Black (20/100)
Red (10/100)
Urn II
Black (90/100)

Urn I (1/10)
Urn II (9/10)

$$\frac{1/10 \text{ X } (80/100)}{1/10 \text{ X } (80/100) + 9/10 \text{ X } (10/100)} = \frac{.04}{.04 + .09} = .31$$

Urn I (2/10)
Urn II (8/10)

$$\frac{2/10 \text{ X } (80/100)}{2/10 \text{ X } (80/100) + 8/10 \text{ X } (10/100)} = \frac{.08}{.08 + .08} = .5$$

Urn I (3/10)
Urn II (7/10)

$$\frac{3/10 \text{ X } (80/100)}{3/10 \text{ X } (80/100) + 7/10 \text{ X } (10/100)} = \frac{.12}{.12 + .19} = .63$$

Urn I (4/10)
Urn II (6/10)

$$\frac{4/10 \text{ X } (80/100)}{4/10 \text{ X } (80/100) + 6/10 \text{ X } (10/100)} = \frac{.16}{.16 + .06} = .73$$

Problem 6

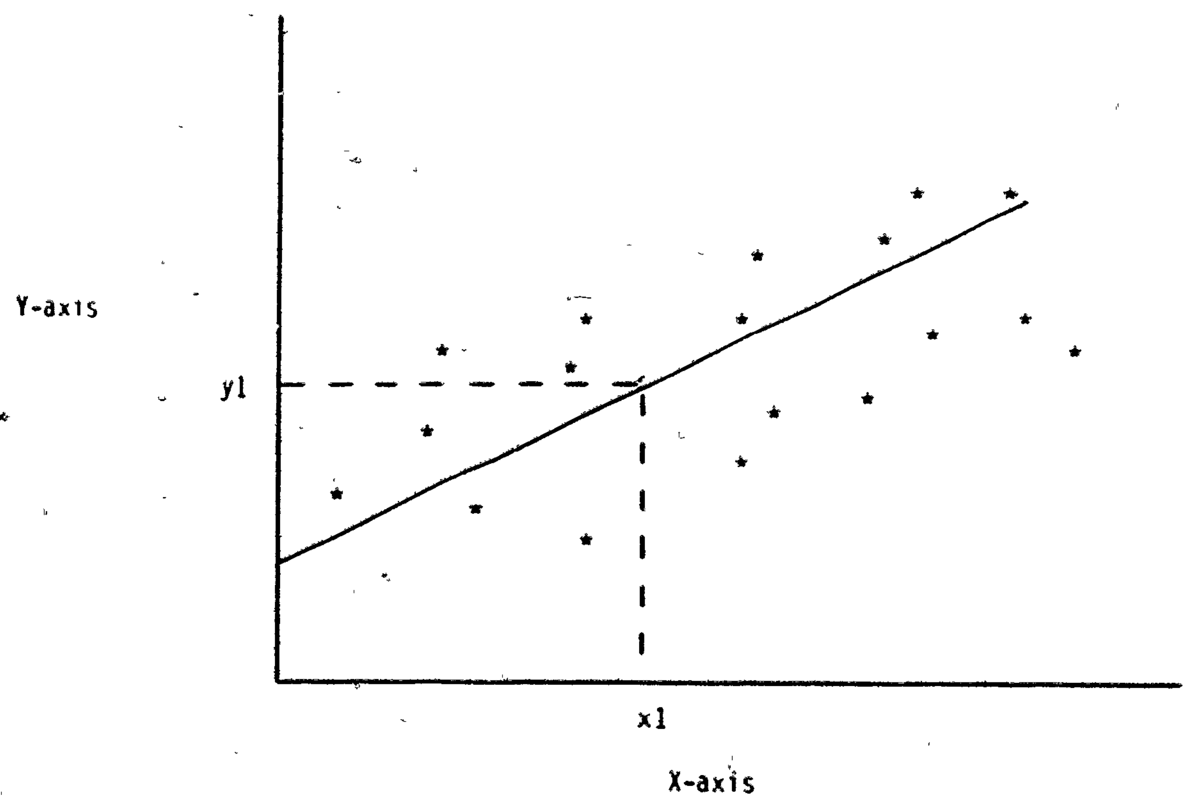
Problem 6 is similar to problem 1. Recall that if you toss a coin 10,000 times you expect, on average, 5,000 Heads and 5,000 Tails. [Show subjects diagram plotting proportion of Heads from problem 1 (Diagram 2)]

With a die, there are 6 possible outcomes, that is, 1, 2, 3, 4, 5, and 6. If you roll a die 6,000 times you expect, on average, 1,000 1's, 1,000 2's ... 1,000 6's. Therefore, you would expect to roll a 6 one in six times (1/6 = .167).

Problem 7

If X and Y are positively correlated, then as values of X increase the corresponding values of Y increase. [Show subjects diagram (Diagram 7)] The points, which represent pairs of X and Y values, spread from lower left to upper right. A numerical index called the correlation coefficient expresses the degree of the relationship between the X and Y values. Using the formula we fit a line through the points. As you can see, the points cluster closely around the plotted line as compared to [move pen on diagram] this line. The problem asks whether you would have more confidence in predicting values of X from values of Y, Y from X, or equal confidence in both. By drawing a straight line up from the X axis at x1 to the plotted line and then drawing a straight line from the point of intersection on the plotted line to the Y axis (y1), you can see that predicting x1 from y1 is the same as predicting y1 from x1.

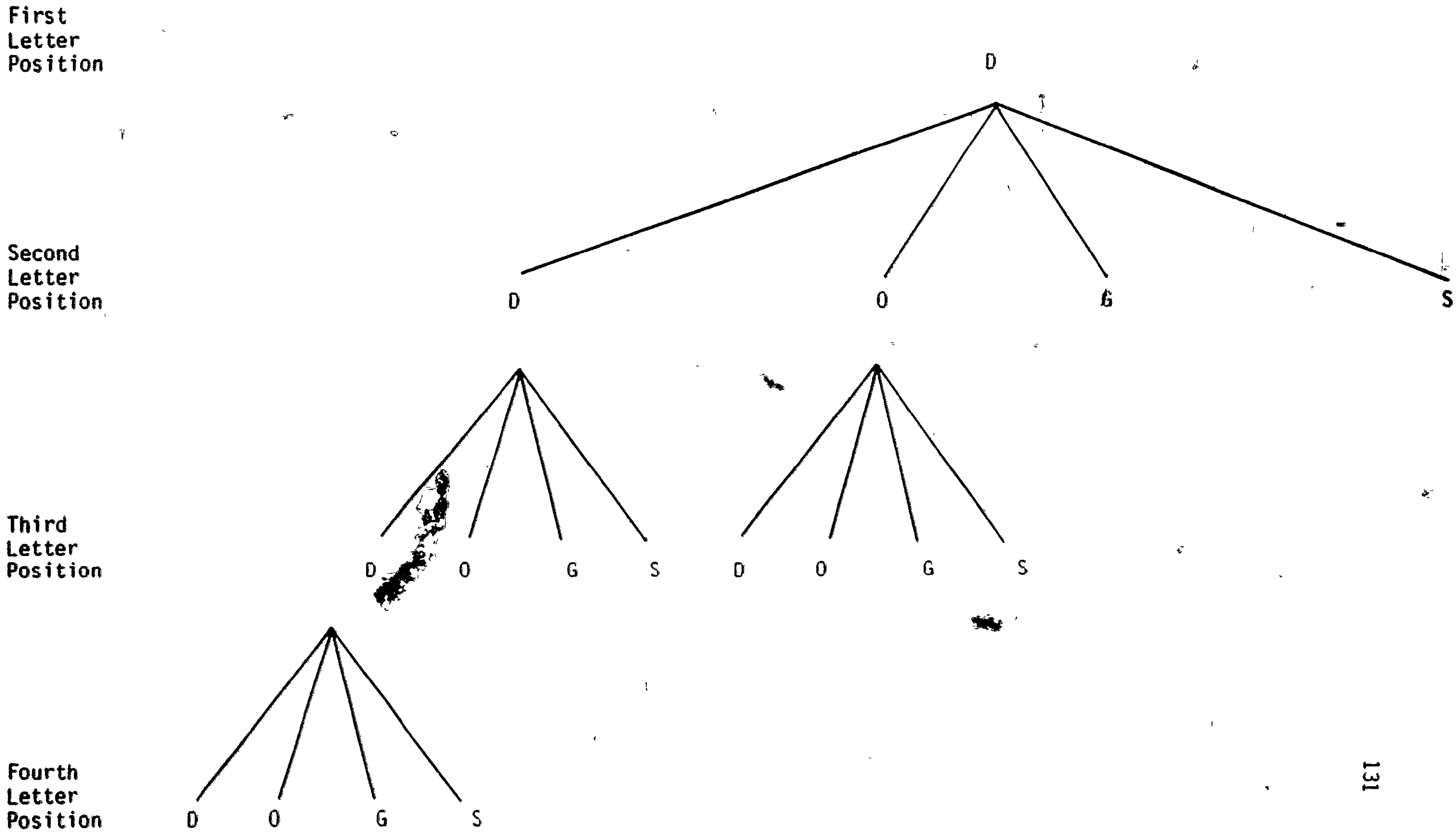
The prediction of X(Y) from Y(X)



Problem 8

The event of interest in problem 8 is the number of sequences, with repetition of letters, one can form with 4 letters. You have the letters D, O, G, and S, and you want to know how many possible 4 letter sequences you can form, with repetition of letters. That is, you can use, say, the letter D in the first, second, third, and/or fourth letter position to form a 4 letter sequence. In the diagram (Diagram 8), each letter position is represented as a compartment, with each compartment containing 4 letters (4 choices of a letter for each position). There are 4 letters that can appear in the first position, 4 in the second position, 4 in the third position, and 4 in the fourth position. [Show subjects first two paths of diagram (Diagram 9)] Therefore, there are 4 X 4 X 4 X 4 = 256 possible 4 letter sequences that can be formed, with repetition of letters, from the word DOGS.

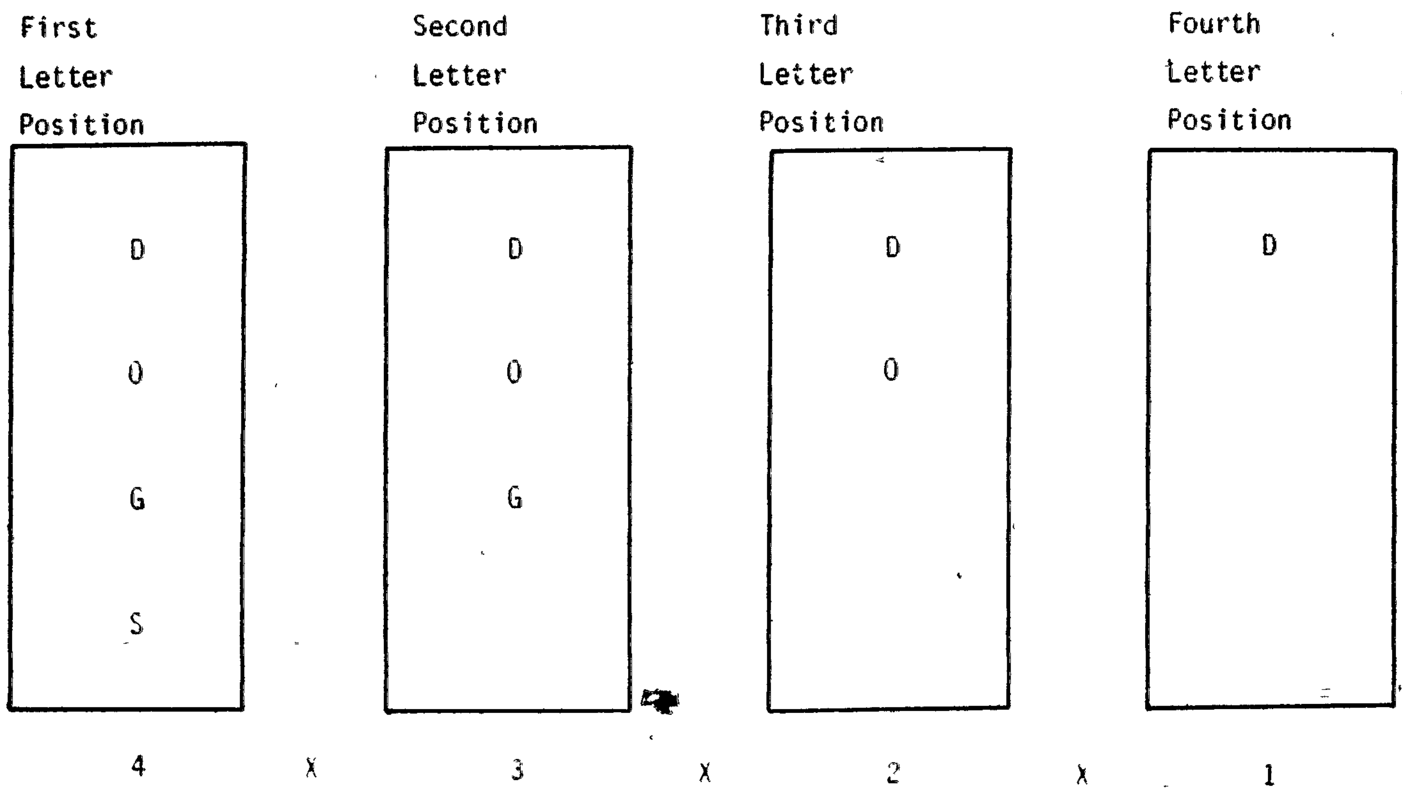
First Letter Position		Second Letter Position		Third Letter Position		Fourth Letter Position
D		D		D		D
O		O		O		O
G		G		G		G
S		S		S		S
4	X	4	X	4	X	4

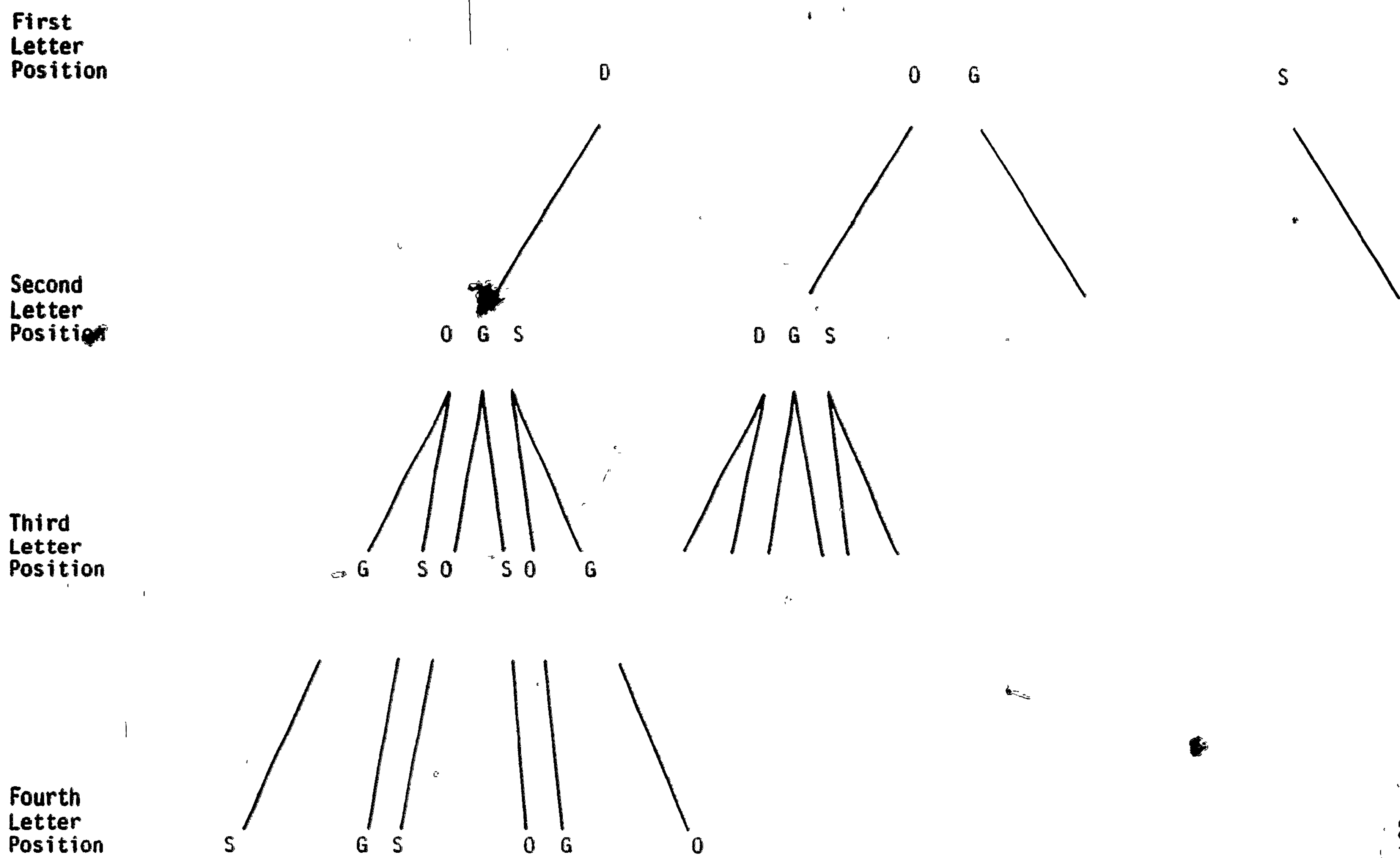


First Letter Position
Second Letter Position
Third Letter Position
Fourth Letter Position
D
D
O
G
S
D
O
G
S
D
O
G
S
D
O
G
S

Problem 9

Problem 9 is similar to problem 8 except that the compartment options decrease by one as you make a choice for each letter position (remember, you may not use a letter more than once in any sequence). In the diagram (Diagram 10), each letter position is represented as a compartment. The number of letter options decreases by one in each compartment. When a letter is chosen for a letter position it cannot be used in any other letter position and is dropped from the subsequent compartment(s) as an option. Therefore, you have 4 X 3 X 2 X 1 = 24 possible 4 letter sequences that can be formed, without repetition of letters, from the word DOGS. The 24 possible sequence are: [Show subjects first two paths of diagram (Diagram 11)]


First Letter Position
Second Letter Position
Third Letter Position
Fourth Letter Position
D
O
G
S
D
O
G
D
O
D
4
X
3
X
2
X
1


First Letter Position
D
O G
S
Second Letter Position
O G S
D G S
Third Letter Position
G S O S O G
Fourth Letter Position
S G S O G O

Problem 10

The event of interest is which jury has the higher probability of making a correct decision. There are three possibilities of making a correct decision for the 3 man jury.

The first possibility is obtained if the 2 members, each of whom independently has a probability p of making a correct decision, both make a correct decision. Majority rules, so the third member who flips a coin for each decision is not included in the decision. Because the decisions of members are independent, we multiply p X p.

1) $p \times p = p^2$

A second possibility of making a correct decision is obtained if the first member makes a correct decision and the second member makes an incorrect decision. Probabilities or proportions fall between 0 and 1. Therefore, if the probability of a correct decision is p then the probability of an incorrect decision is 1 - p. The third member who flips a coin for each decision will be included in the decision. The third member will, on average, make a correct decision 1/2 of the time

2) $p \times (1-p) \times 1/2 = 1/2p - 1/2p^2$

A third possibility of making a correct decision is obtained if the first member makes an incorrect decision and the second member makes a correct decision. The third member, who flips a coin for each decision will be included in this decision.

3) $(1-p) \times p \times 1/2 = 1/2p - 1/2p^2$

Therefore, there are three ways the 3 man jury can make a correct decision, that is, 1) or 2) or 3). The total possibilities of making a correct decision is obtained by adding 1) + 2) + 3) to obtain:

$$1/2p - 1/2p^2 + (1/2p - 1/2p^2) + p^2$$

$$= p - p^2 + p^2$$

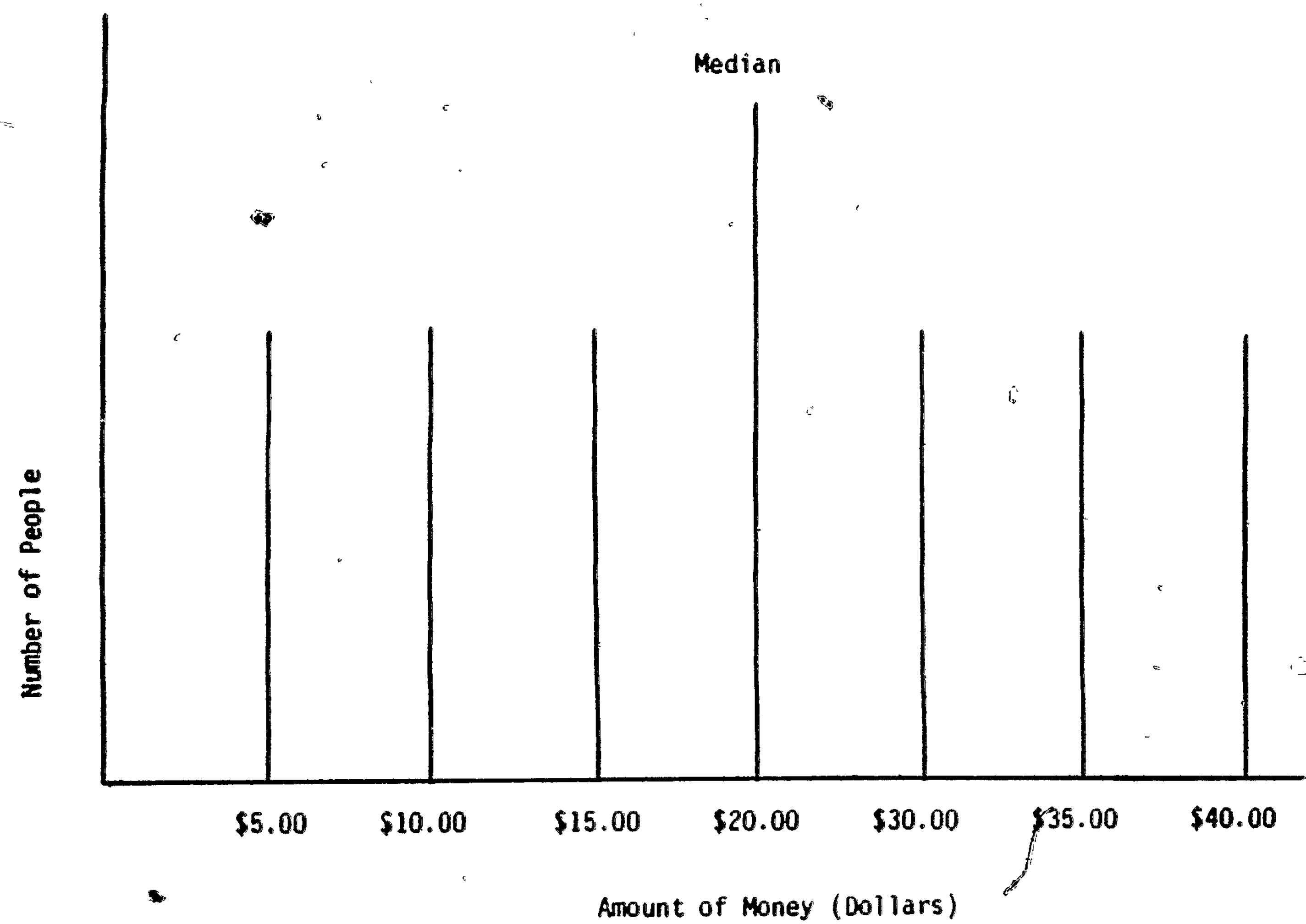
$$= p$$

The 1 man jury also has probability p of making the correct decision. Therefore, the 3 man jury is equally likely to make a correct decision.

Additional Problem

The event of interest is whether Method 1 or Method 2 will have more people with cash greater than $30.00. All we know is that the number of people who carry more than $20.00 is equal to the number of people who carry less than $20.00. [Show subjects distribution of cash (Diagram 12)] There are three amounts less than $20.00 ($5.00, $10.00, and $15.00) and three amounts greater than $20.00 ($30.00, $35.00, and $40.00). If you were to put these amounts on separate pieces of paper and draw either one piece of paper and record the amount (Method 1) or three pieces of paper and record the middle amount (Method 2), you would find, on average, that it is easier to draw one amount and have it greater than $30.00 than it is to draw three amounts and have two greater than $30.00. The variance will be larger in Method 1 than in Method 2. If we were to restrict ourselves to the 7 dollar values shown on the hypothetical distribution of cash, you can see that the probability of selecting a dollar value greater than $30.00 using Method 1 is 2/7. With Method 2, the probability of selecting 2 dollar values greater than $30.00 is 2/7 X 1/6 = 2/42.

Hypothetical Distribution of Cash



Appendix E

ANOVA Table and Means and Standard Deviations for Probability Knowledge Questionnaire Scores

ANOVA Summary Table for Probability Knowledge Questionnaire Scores

SOURCE	SS	DF	MS	F	P
G&C-O	159.612	1	159.612	169.612	.000
T&K	2.112	1	2.112	2.241	.139
G&C-O x T&K	2.112	1	2.112	2.241	.139
RESIDUAL	71.650	76	.943		
TOTAL	235.487	79	2.981		

Note. GEBOTYS AND CLAXTON-OLDFIELD classification (G&C-O). TVERSKY and KAHNEMAN classification (T&K)

Mean Probability Knowledge Questionnaire Scores as a Function of Classification and Training

Classification: Knowledge - Education		Condition: Training	Condition: No Training
Novice - Novice			
	M	4.20	3.60
	SD	1.03	0.97
	N	10	10
Novice - Expert			
	M	3.90	3.80
	SD	0.88	0.79
	N	10	10
Expert - Novice			
	M	6.20	6.50
	SD	0.63	0.71
	N	10	10
Expert - Expert			
	M	7.00	7.00
	SD	1.15	.49
	N	10	10

Note. Maximum score = 10

Appendix F

ANOVA Table and Means and Standard Deviations for Probability Test Scores

ANOVA Summary Table for Probability Test Scores

SOURCE	SS	DF	MS	F	P
G&C-0	61.250	1	61.250	27.189	.000
T&K	11.250	1	11.250	4.994	.029
TRN	76.050	1	76.050	33.758	.000
G&C-0 x T&K	7.200	1	7.200	3.196	.078
G&C-0 x TRN	.800	1	.800	.355	.553
T&K x TRN	9.800	1	9.800	4.350	.041
G&C-0 x K&T x TRN	2.450	1	2.450	1.088	.301
RESIDUAL	162.200	72	2.253		
TOTAL	331.000	79	4.190		

Note. GEBOTYS and CLAXTON-OLDFIELD classification (G&C-0). TVERSKY and KAHNEMAN classification (T&K). Training (TRN)

Mean Probability Test Scores as a Function of Classification and Training

Classification: Knowledge - Education		Condition: Training	Condition: No Training
Novice - Novice			
	M	5.90	2.70
	SD	1.66	0.95
	N	10	10
Novice - Expert			
	M	5.00	3.90
	SD	0.82	0.88
	N	10	10
Expert - Novice			
	M	6.50	4.40
	SD	1.65	1.51
	N	10	10
Expert - Expert			
	M	7.50	6.10
	SD	1.27	2.51
	N	10	10

Note. Maximum score = 10

Appendix G

Proportions of Subjects Correctly Answering the Probability Test Problems

Proportion of Correct Answers to the Probability Test Problems as a Function of Training

	Condition	
Problem	Training	No Training
1	.55	.52
2	.75	.20 *
3	.90	.65 *
4	1.00	.82 *
5	.10	.13
6	1.00	.92
7	.95	.67 *
8	.42	.17 *
9	.32	.13 *
10	.22	.05 *

Note. $*p < .05$