Journal of Experimental Psychology:
Human Perception and Performance
1985, Vol. 11, No. 4, 443-456

Copyright 1985 by the American Psychological Association, Inc.
0096-1523/85/$00.75

Revision of Beliefs When a Hypothesis Is Eliminated From Consideration

Lori B. Robinson and Reid Hastie
Northwestern University

In three experiments, strengths of belief in several competing hypotheses were examined after the elimination of hypotheses from the original set under consideration. Subjects read murder mysteries, each containing a series of clues pertaining to a closed set of suspects. After each clue, subjects estimated the probability of guilt for each suspect. Some subjects were tutored to revise their judgments in accordance with probability theory, adjusting probabilities in such a way that they maintained a sum of 1.0 across the set of hypotheses. Untutored subjects, however, were poor intuitive probability theorists. When a suspect was eliminated, they tended not to revise probabilities for remaining suspects. Furthermore, following any given clue they tended to adjust only the probability of one target suspect, resulting in a constantly changing sum of the probabilities. Subjects seemed to be using an anchor-and-adjust strategy in revising their beliefs but treated each hypothesis as if it were independent of its alternatives.

Uncertainty about the world forces us constantly to consider new evidence and to revise our opinions regarding our beliefs. Although much research has concentrated on the impact of evidence on a stable set of hypotheses, little has been said about possible strategies for dealing with a changing set of hypotheses. Our own curiosity about the process of opinion revision was recently piqued when we asked some of our collegues what they thought might happen in judging several competing hypotheses when one hypothesis was suddenly eliminated from consideration. Logically we would expect all competing hypotheses to increase in probability under these circumstances. Bayes's theorem, which is easily applied to our opinion revision question, prescribes revision of all prior probabilities in proportion to their strengths: The likeliest prior probability would gain the most; the least likely would gain the least.

Despite the simplicity of the Bayesian calculation, our colleagues' intuitions favored several alternate revision patterns. After all, Bayesian formulations do not usually model individual subjects' judgment patterns (Edwards, 1968; Fischhoff & Lichtenstein, 1978; Tversky & Kahneman, 1974). The more mathematically minded subject might reason that because one hypothesis with a given probability has been eliminated, all other hypotheses under consideration should gain in probability. However, this gain could be distributed evenly among the remaining hypotheses rather than in proportion to the preelimination probabilities, as Bayes's theorem prescribes. Alternatively, subjects might opt for a simpler type of revision, such as increasing the probability of only one of the remaining hypotheses (e.g., the current favorite).

Changes in a subject's confidence could lead to more complicated revisions. If an elimination increases a subject's confidence in his or her current opinion, the result could be a spreading effect involving increased probabilities for favored hypotheses and decreased probabilities for the long shots. Conversely, a subject who loses confidence after an elimination might increase long shot probabilities and decrease probabilities of the current favorites. Even more drastic reorderings could also occur; the elimination of a favored hypothesis might conceivably lead a subject to select new dimensions of the evi-

This research was supported in part by National Science Foundation Grant SES-8208132.

The authors are grateful to Geoffrey Fong, Peter Frey, and Nancy Pennington for useful advice on the research.

Requests for reprints should be sent to Reid Hastie, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

dence on which to focus and thereby create a completely new ranking of the remaining hypotheses.

Thus, we found ourselves with a list of plausible conjectures concerning postelimination hypothesis revision behavior. Empirical research was the obvious next step. We ran a series of three studies in which subjects were given a short mystery story and a series of clues; after each clue, subjects were asked to estimate the probability of guilt for each suspect. At various points in the clue presentation, evidence that completely eliminated a suspect from further consideration was presented.

We hoped to gain insight into the processes used by subjects as they revised their probability estimates, both before and after the elimination of a hypothesis. In the first study, we compared subjects' performance with mathematical norms in both tutored and untutored situations. In the second study, we looked at two characteristics of the eliminated hypotheses that might affect subject behavior in dealing with an elimination. Finally, in the third study, we investigated in greater detail the effect of particular combinations of evidence on the subjects' judgments.

Throughout this article, we will refer to subjects' ratings as *probabilities.* This term is not meant to imply that such ratings do (or should) behave according to the laws of mathematical probability; instead, the term should be read as strength of belief in a hypothesis. In operational terms, in the research reported in this article, this strength is measured by subjects' estimates expressed on a 5-in. (12.7-cm) rating scale labeled 0.0 to 1.0 as a probability continuum.

Experiment 1

Past research (e.g., Edwards, 1968; Phillips & Edwards, 1966; Phillips, Edwards, & Hays, 1966) has indicated that subjects' beliefs do not usually conform to the rules of probability, such as Bayes's theorem, in opinion revision tasks. It might be anticipated that in our task, too, subjects would exhibit behavior deviating from mathematical norms. A natural question would then be, why do subjects not follow the rules? One way of gaining insight into performance deficiencies is to tutor some subjects in probability theory. We used this technique in the hope of discovering (a) whether subjects could be induced through training to follow mathematical probability rules in evaluating competing hypotheses and (b) whether subjects not specifically trained in using these rules would naturally follow them.

Recent research (Fong, Krantz, & Nisbett, 1983; Nisbett, Krantz, Jepson, & Kunda, 1983) has suggested that one reason subjects fail to use probability laws in making judgments is that those subjects simply may not recognize the probabilistic nature of the task. Solving murder mysteries might be a task that subjects would not naturally categorize as probabilistic. Our tutoring condition, then, consisted of an explanation of the relevant probability rules and explicit instructions in their applicability to the task at hand. Two groups of subjects, one tutored and one untutored, both attempted to solve two mystery stories. As clues were presented, subjects were asked to rate each suspect's likelihood of guilt. Toward the end of the series of clues, two suspects were eliminated from further consideration, and new ratings were made after each of these eliminations. We wished to know if subjects in the tutored group would show greater tendencies to adjust their ratings for the remaining suspects after an elimination and if their ratings would in general conform to the laws of probability better than their untutored counterparts' ratings. Improved performance by the tutored group would indicate the extent to which nonprobabilistic thinking could be attributed to simple lack of recognition of the appropriateness of a probabilistic representation of the task.

It was also anticipated, however, that conformity to the rules of probability might overload subjects' processing capacities and thus lead tutored subjects to adopt simplifying strategies in their approach to the task. For example, Wright (1974) has shown that subjects operating under a high information load in a choice situation may simplify their task by attending to fewer data dimensions. Our task, particularly the tutored condition, might lead to similar judgmental shortcuts, such as consideration of fewer alternative hypotheses at any given time. (See also Levine, 1970,

and others on consideration of a limited number of hypotheses in concept attainment tasks.)

Method

Subjects. Subjects were 40 male and female undergraduate students at Northwestern University who participated in the experiment in order to fulfill a course requirement. Subjects were run individually in single sessions lasting approximately 90 min.

Materials. Two mystery stories, "The Murdered Banker" and "The Poisoned Philanthropist," were written for the present experiment. These stories were each composed of two parts: a brief (380–640 words) plot scenario, which set the scene and introduced the victim and five suspects, and a set of 13 clues. These clues were of four types: (a) 3 guilty clues, which provided information implicating one particular suspect; (b) 3 innocent clues, which pointed toward the innocence of a suspect; (c) 5 neutral clues, which provided information about the victim or the crime itself that did not directly relate to any one suspect; and (d) 2 eliminators, which gave a suspect an airtight alibi and thus eliminated him or her from further consideration. Clues were designed to satisfy conditional independence; that is, each clue referenced only one suspect, and elimination of one suspect did not logically point toward any one remaining suspect more than the others. The eliminators were presented late in the series as the 11th and 13th clues. All clues were pretested on a small group of graduate and undergraduate students to insure that the clues would be interpreted as we had expected.

Two different sets of training materials were constructed for the two groups of subjects. The instructions for the tutored group included training in the permissible range of probabilities (0.0 to 1.0) and in the additivity of probabilities for mutually exclusive and collectively exhaustive events. These materials also included a set of practice exercises involving the estimation of probabilities and the additivity of probabilities for such exclusive, exhaustive sets of hypotheses. Some of these practice exercises explicitly involved estimating probabilities of guilt for suspects in murder mysteries. Training materials for the untutored group included only instruction on the range of probabilities and a set of five sample probabilities (none of which involved mysteries) to estimate for practice.

Subjects' ratings were made on 5-in. (12.7-cm) calibrated scales running from 0.0 (labeled *no chance*) to 1.0 (labeled *sure thing*), marked off in intervals of .10. Rating scales for each suspect and each clue were printed on separate sheets of paper.

Design and procedure. Subjects were informed that they were participating in an experiment on human judgment and that the judgments they would be asked to make involved suspects in two different murder mysteries. Training materials were read to all subjects, with 20 subjects receiving the instructions for the tutored group and 20 subjects receiving the instructions for the untutored group. Every subject attempted to solve both "The Murdered Banker" and "The Poisoned Philanthropist"; the order of case presentation was counterbalanced within both the tutored and untutored groups.

For each story, the experimenter read the plot scenario to the subjects once. To insure comprehension, subjects were then given a recall test that consisted of six questions involving the identification of suspects, the time and place of the murder, and alibis of selected suspects. The plot scenario was then read a second time and the recall test given again. Subjects were told that the guilty party would definitely be one of the five suspects introduced in the scenario and that there would be no conspiracy among two or more suspects; the murderer would always act alone.

Subjects were then given separate rating scales for each suspect and were instructed to make a slash through the scale at the point that they thought corresponded to the probability that the suspect was guilty of the murder. Subjects were also asked to think aloud as they made these ratings and all subsequent ratings.

After these baseline ratings were made for all five suspects, the experimenter read the first clue to the subjects and asked them to once again rate each suspect's probability of guilt. The rating scales were scrambled so that suspects were considered in a new order on each trial. This procedure was repeated for all 13 clues, at the end of which the experimenter announced the solution to the mystery. The above procedures were then repeated for the second case.

Results

Preelimination ratings. For both groups, ratings for each suspect after each clue were averaged, and these mean ratings were graphed for both cases. (Although the dangers of averaging such data are recognized, in this case averages provide a useful summary of general trends in the behavior of individual subjects.) These graphs showed a strong tendency for untutored subjects' ratings to remain stable for a given suspect except after a clue that specifically referred to that suspect. A guilty clue referring to George Robner in "The Poisoned Philanthropist," for example, typically produced an increase in George's probability rating but no change in the ratings of the other four suspects. The tutored subjects, however, tended to adjust their ratings for all five suspects on every guilty or innocent clue (see Figures 1 and 2).

For both groups and both cases, averages were also computed for the total probability (sum across five ratings) assigned to the five suspects after each clue. These summed ratings tended to remain stable at 1.00 for the tutored subjects, in accordance with probability rules. Summed ratings for untutored subjects, however, took on values ranging from 1.39 to 2.32 and tended to increase or decrease depending on the nature of the clue presented (see Figure 3).

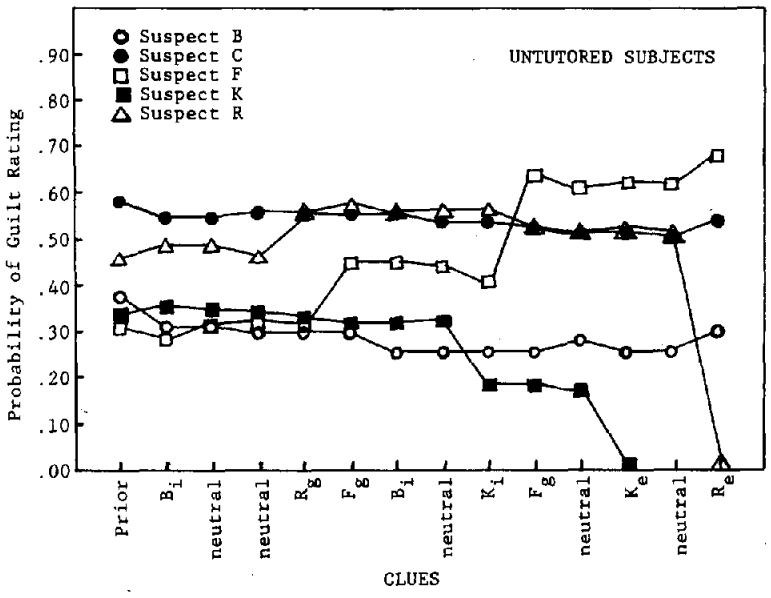

Figure 1. Mean probability of guilt ratings for untutored subjects, Experiment 1, "The Murdered Banker." (Clues are presented on the ordinate, with capital letters indicating the target suspect [B, C, F, K, R] and subscripts indicating the valence of the clue [g = guilty, i = innocent, e = eliminator]. Prior probabilities were estimated after presentation of the plot scenario but before exposure to clues.)

It had been hypothesized that subjects in the tutored group might find conforming to the laws of probability to be an unusually difficult task. One plausible strategy for simplifying the task would be to narrow down the set of suspects. To test whether this occurred, the average number of suspects under consideration after each clue (i.e., rated

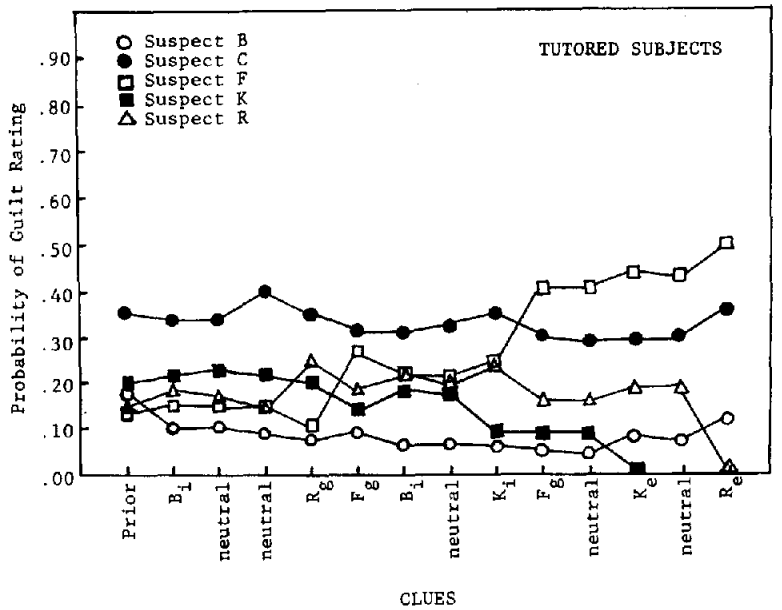

Figure 2. Mean probability of guilt ratings for tutored subjects, Experiment 1, "The Murdered Banker."

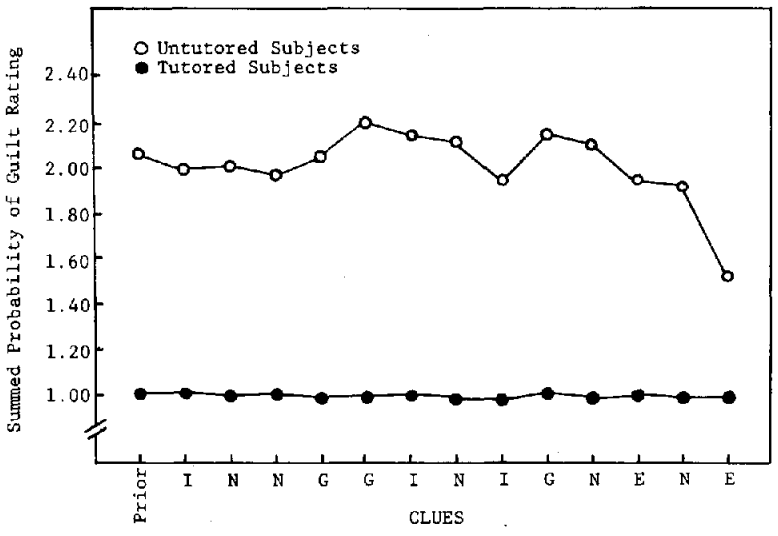

Figure 3. Mean summed probability of guilt ratings for tutored and untutored subjects, Experiment 1, averaged across both cases. (Clues are presented on the ordinate, with clue valence indicated as follows: G = guilty clue, I = innocent clue, N = neutral clue, E = eliminator.)

above 0.0) was calculated and graphed for both groups of subjects. The number of suspects under consideration was consistently higher for untutored subjects than for tutored subjects (see Figure 4). Zero ratings did not seem to result from subjects' rounding down their smaller probabilities; most subjects tended to make use of very low ratings (e.g.,

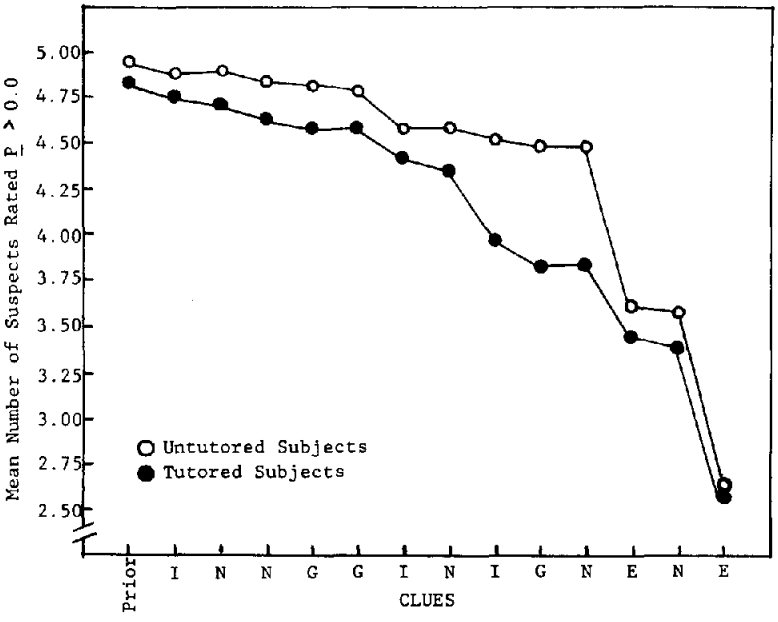

Figure 4. Mean number of suspects under consideration for tutored and untutored subjects, Experiment 1, averaged across both cases.

.02) as distinct from zero ratings, which were often accompanied by such comments as "Well, this suspect is out of it."

Conventions of mathematical probability theory tell us that once a suspect has been rated at 0.0 probability, that suspect is permanently eliminated; he or she can never return to a nonzero rating no matter how strong the evidence. Schum and Martin (1980), however, found that subjects often moved a zero-rated hypothesis back to a positive rating; they called these phenomena *resuscitations.* Our subjects, too, exhibited this behavior. Tutored subjects resuscitated suspects in 30 out of the 65 instances in which a suspect had been rated at 0.0. Of these resuscitations, 7 occurred immediately after a guilty clue that implicated the zero-rated suspect, 6 occurred after an innocent clue that weakened the case against a more favored suspect, and 6 occurred after an elimination. The remainder seemed to occur haphazardly, although 7 occurred after neutral clues, which may have been interpreted as implicating one suspect or another. Untutored subjects, who rated far fewer suspects at 0.0, made fewer resuscitations, although their proportion of resuscitations (10 out of 29 opportunities) was still high. Three of these resuscitations occurred after eliminations, 1 after an innocent clue, and the remainder after neutral clues.

A detailed analysis of individual (rather than averaged) data was also conducted. For subjects in both groups, ratings after each clue were categorized in terms of (a) direction of adjustment in ratings for the target suspect (that suspect specifically referred to in the clue) and (b) direction of the mean adjustment to the four nontarget suspects. Adjustment to the target suspect was generally in the desired direction for both groups of subjects. Guilty clues resulted in increases in the target's rating 91% of the time for tutored subjects and 83% of the time for untutored subjects. Innocent clues were somewhat less successful, resulting in 61% decreases to the target for the tutored group and 63% decreases for the untutored group. Neutral clues, which had no target, resulted in no change in ratings in 62% of tutored subjects' judgments but only 31% of untutored subjects' judgments. Of greater interest, however, are the average adjustments to nontarget suspects, which should have been made in a direction opposite that of the target adjustment according to the rules of probability. Tutored subjects conformed to these rules on 88% of all guilty clues and 53% of all innocent clues. Untutored subjects, on the other hand, conformed to these rules on only 47% of guilty clues and 31% of innocent clues (the latter percentage representing only chance performance). Untutored subjects were much more likely than tutored subjects to make no adjustment to nontarget suspects. Differences between the two groups were statistically significant, $\chi^2(8, N = 880) = 324.23$, $p < .001$. There was no significant difference between the two sets of case materials, "The Murdered Banker" and "The Poisoned Philanthropist," in terms of these response categories (see Table 1).

Table 1
Preelimination Adjustment Strategy Usage as a Function of Clue Type and Training Condition[a]

Condition	Clue type		
	Guilty	Innocent	Neutral
Tutored subjects			
No change in nontarget ratings	11	33	62
M probability increases	1	53	19
M probability decreases	88	14	19
Untutored subjects			
No change in nontarget ratings	36	34	31
M probability increases	17	31	36
M probability decreases	47	35	33

Note. Numbers are percentages. *n*s = 120 for guilty and innocent and 200 for neutral.
[a] Averaged across both cases.

Postelimination ratings. Individual subjects' responses to an elimination were broken down into eight exclusive strategies: (a) flat (subject does not adjust probabilities at all for remaining suspects after an elimination); (b) add to all (subject adjusts the probabilities of all remaining suspects upward, but amounts added to each are not necessarily proportional to preelimination probabilities);

(c) add to favorites (subject adjusts the probability of only the highest rated suspect(s) upward); (d) add to middle (subject adjusts the probability of medium-likelihood suspect(s) upward); (e) add to long shots (subject adjusts the probability of only the lowest rated suspect(s) upward); (f) spreading (subject adds to the probability of high-rated suspects and subtracts from the probability of low-rated suspects); (g) narrowing (subject adds to the probability of low-rated suspects and subtracts from the probability of high-rated suspects); and (h) miscellaneous (not classifiable as any of the previous seven strategies). The number of suspects in each group using a particular strategy was counted for both eliminations in both cases. Untutored subjects overwhelmingly preferred the flat strategy on the first elimination (20 out of 40 fell into this category, with the remainder evenly split among the other seven strategies) but were likely to use some form of adjustment, particularly the various adding strategies or the spreading strategy, on the second elimination. Tutored subjects also favored the flat strategy for the first elimination, though not to such a degree (9 out of 40, with the remainder evenly split among adjustment strategies). On the second elimination, tutored subjects favored add-to-all (11), add-to-favorites (12), and add-to-long-shots (10) strategies. Differences between the two groups were statistically significant, $\chi^2(7, N = 160) = 45.86$, $p < .001$, as were the differences between the two eliminations, $\chi^2(7, N = 160) = 86.06$, $p < .001$. There was no significant difference between the two cases in terms of favored strategies.

To compare subjects' ratings quantitatively to the optimal mathematical probabilities, Bayes's theorem was used to calculate post-elimination adjustments for each individual subject. The numbers of accurate Bayesian revisions (within $\pm.02$ of the Bayesian estimate), overestimations, and underestimations were counted for each group. Slightly over half (58%) of the tutored subjects' ratings on the first elimination were Bayesian, and the remainder were evenly split between over and underestimations. (The data showed tendencies toward overestimation of very low probabilities and underestimation of higher probabilities.) However, many of these Bayesian adjustments were made when the eliminated hypothesis had already been rated at zero by the subject, that is, when even Bayes's theorem prescribed no adjustment. If we do not consider these special cases, the rate of agreement with Bayes's theorem falls to 39%. On the second elimination, tutored subjects' ratings were Bayesian 36% of the time, again with equal numbers of over- and underestimations. Untutored subjects showed strong tendencies to revise their ratings less than Bayes's theorem prescribed; 58% of their first-elimination ratings and 76% of their second-elimination ratings underestimated the Bayesian value.

Discussion

With proper training, subjects can be induced to follow probability rules to some extent. They can be taught to assign probabilities to mutually exclusive and collectively exhaustive hypotheses so that these probabilities sum to 1.0, and they can subsequently revise those probabilities so that they continue to sum to 1.0 as new evidence is acquired. Complementarity of probability ratings is achieved, but adjustments to nontarget ratings are still not necessarily Bayesian. Thus, recognition of the relevance of probability rules does seem to improve subjects' performance in a limited sense, but there is little evidence of any gain in insight into probabilities and their mathematical operations.

Conformity to even the simple additivity rule, however, does not come easily to tutored subjects. Their verbal reports indicate that the task is difficult; further, their behavior exhibits signs of their attempts to simplify the task. These subjects begin eliminating hypotheses (narrowing the set of alternatives) as early information comes in and keep consistently fewer hypotheses under consideration than do subjects left to their own devices. However, even tutored subjects do not seem to regard their eliminations as permanent. When new evidence is acquired, hypotheses rated at a zero probability are frequently revised to a nonzero probability. In commonsense terms, such resuscitations may be desirable, in that they indicate a subject's willingness to reconsider eliminations that may have been based on too little evidence. From a Bayesian standpoint, however, when a hy-

pothesis is rated at zero, it should remain at zero.

When subjects are not given specific training in the laws of probability and their implications for a judgment task, they tend to be poor intuitive probability theorists. Even when told that a set of hypotheses is mutually exclusive and exhaustive, subjects give ratings that sum to more than 1.0. This failure to sum to 1.0 appears to be more than simply a scaling problem or a quantitative failure to normalize probability ratings; the tendency of subjects to make no revisions after the elimination of a hypothesis is strong evidence of a qualitative failure to appreciate the complementarity of probabilities. When adjusting the probability of any hypothesis upward or downward, subjects tend not to make the corresponding adjustments to other hypotheses in the set.

Subjects' non-Bayesian treatment of eliminations, however, is dependent on the particular elimination in question. As seen in their patterns of elimination strategies, our subjects typically made no adjustments after the first elimination but were more likely to make some adjustment after the second elimination. There are two plausible explanations for this difference: (a) The order of elimination may be the critical variable in determining whether adjustments will be made or (b) because, for both sets of case materials, the second elimination involved a higher probability suspect than the first, the critical variable may be the preelimination probability of the eliminated hypothesis. Experiment 2 attempted to test both of these explanations of the data.

Experiment 2

This study was designed to vary the preelimination probability of an eliminated suspect. The number of guilty clues implicating the suspect who is later the first to be eliminated was varied between zero and four. It was hypothesized that this first elimination would induce little or no adjustment of probability ratings when there were no clues against the eliminated suspect (i.e., when the eliminated suspect was a low-probability suspect) but would lead to adjusted ratings when there were four clues against the eliminated suspect (i.e., when it was a high-probability suspect). This study also endeavored to look more closely at the effects of order of elimination; four eliminations were used rather than the two of Experiment 1. It was hypothesized that later eliminations would lead to greater adjustments in probability than would early eliminations.

Method

Subjects. The subjects were 60 male and female Northwestern undergraduates enrolled in a general psychology course. Subjects were run individually in single sessions lasting approximately 45 min.

Materials. Training materials were identical to those used for the untutored group in Experiment 1.

The two plot scenarios from Experiment 1, "The Murdered Banker" and "The Poisoned Philanthropist," were again used. Clue sets for each story were now expanded to include a total of 22 clues. Three different sets of clues were constructed for each story. In the Condition 1 set, Clues 4, 8, 11, and 16 were neutral. For Condition 2, Clues 4 and 11 were neutral but Clues 8 and 16 were guilty clues implicating the suspect who is later the first to be eliminated. For Condition 3, Clues 4, 8, 11, and 16 all were guilty clues implicating the first eliminated suspect. The remainder of the clues were the same for all three conditions; these included 4 eliminators (Clues 17, 19, 21, and 22); 2 neutral clues (18 and 20); 2 guilty clues and 1 innocent clue referring to the second eliminated suspect; 3 guilty clues and 1 innocent clue referring to the third eliminated suspect; 1 guilty clue and 1 innocent clue referring to the fourth eliminated suspect; and 2 guilty clues referring to the murderer.

Rating scales were again printed on separate sheets of paper for each suspect and each clue, as per Experiment 1.

Design and procedure. Each subject attempted to solve only one case; 30 subjects received "The Murdered Banker" and 30 subjects received "The Poisoned Philanthropist." Within each group of 30, 10 subjects received the Condition 1 (low-probability eliminated suspect) clue set, 10 received Condition 2 (medium-probability), and 10 received Condition 3 (high-probability).

The second experiment was conducted following the same procedures as the first, with one change: Subjects did not think aloud as they made their ratings.

Results

Preelimination ratings. For each of the six groups, ratings for each suspect after each clue were averaged, as in Experiment 1. Again, for all groups, subjects' ratings for a given suspect tended to remain relatively stable except after a clue directly referring to that suspect. Averages were also computed for the total probability assigned to the five suspects after each clue. These summed probabilities took on values consistently higher than 1.0 (peaking at 2.54) and tended to increase or

decrease depending on the nature of the clue presented.

Individual subjects' ratings after each clue were again categorized in terms of adjustment to target and nontarget suspects, as per Experiment 1. The percentage of responses that qualitatively conformed to the rules of probability varied from a low of 23% (neutral clues) to a high of 49% (guilty clues). All groups of subjects showed strong tendencies to make no compensatory adjustments to nontarget suspects when increasing or decreasing the rating of a target.

Postelimination ratings. Strategies used by individual subjects after an elimination were categorized as in Experiment 1. The number of subjects using a particular strategy was counted and tabulated as a function of (a) order of elimination and (b) preelimination probability of the first eliminated suspect (Conditions 1, 2, and 3). Subjects ($N = 60$) strongly preferred the flat strategy (making no adjustments 35% of the time) on the first elimination but were equally divided between flat (17%), add-to-all (20%), and add-to-favorites (20%) strategies on the second elimination. By the third elimination, subjects preferred the add-to-all strategy (42%), with only 18% flat strategies used. Differences between the three eliminations were significant, $\chi^2(7, N = 180) = 33.39$, $p < .001$. Subjects in Conditions 1 and 2 (low- or medium-probability eliminated suspect) showed a slightly greater tendency to use the flat strategy (40% for each of these groups, $ns = 20$) than did subjects in Condition 3 (high-probability eliminated suspect, 25% flat strategies), but these differences were not statistically significant.

As a second and possibly stronger measure of the effect of prior probabilities, we looked at strategies used in situations where a subject's least likely suspect was eliminated ($n = 56$) and situations where a subject's most likely suspect was eliminated ($n = 45$), across all three eliminations. Subjects were much more likely to use a flat strategy (36%) when the least likely suspect was eliminated and more likely to use the add-to-all (51%) or add-to-favorites (24%) strategies, with only 11% flat strategies, when the most likely suspect was eliminated. These differences were significant, $\chi^2(7, N = 101) = 34.15$, $p < .001$.

Discussion

As in Experiment 1, we see that subjects in general do not conform to the rules of probability in making and revising probability estimates. Summed probabilities for mutually exclusive and exhaustive hypotheses again tended to equal more than 1.0 and tended to increase or decrease as new evidence was acquired. Resuscitations of zero-rated suspects were also common. When subjects increased or decreased the probability rating of a target suspect, they did not necessarily adjust the probabilities of the other suspects in a complementary fashion. They often made no adjustments at all to nontarget suspects or even adjusted them in the wrong direction.

Subjects' preferred strategies for dealing with the elimination of a suspect depended on the order of elimination and also on the preelimination probability of the eliminated suspect. Subjects tended to prefer a flat strategy, making no adjustments to the ratings of remaining suspects, on early eliminations and eliminations of low-probability suspects. Subjects were more likely to adjust their ratings in some way (primarily by adding to all or adding to the favorites) on later eliminations and eliminations of high-probability suspects. Thus, in qualitative terms, subjects tended to conform to the laws of probability more when more than one elimination had occurred or when the current elimination involved a relatively likely suspect.

The effects of order of elimination pose another interesting problem. It seems that the greater the number of eliminations that have occurred, the less likely a subject is to use a flat strategy. This phenomenon, however, suggests at least two possible explanations. It may be that some sort of learning is taking place as eliminations occur, modifying the subject's approach to later eliminations as they are encountered. A second possibility, though, is that the critical factor is not the number of eliminations that have occurred in the past but the number of hypotheses that remain after a particular elimination. According to the latter explanation, a larger set of remaining hypotheses (i.e., a larger information load) could trigger the flat strategy—a more simple revision than the adjustment strategies. Which of these accounts is

correct is a matter for future research. We are inclined to favor the second explanation at this point for two reasons: (a) Learning seems improbable in the absence of feedback, and (b) a learning process should have caused different patterns of behavior for the first and second cases presented in Experiment 1, but such patterns were not observed.

Experiment 3

The previous two experiments have both involved mysteries with five suspects; it would be valuable to know if the behaviors seen under these conditions will generalize to situations involving a greater number of hypotheses. Thus, in this experiment we constructed two new versions of "The Murdered Banker" and "The Poisoned Philanthropist," this time including nine suspects for each case. We again used a series of three eliminations, this time involving one low-, one medium-, and one high-probability eliminated suspect. We varied the order in which these eliminations occurred to better study any interactions that might occur between order-of-elimination and probability-of-eliminated-suspects effects.

To gain a better understanding of the type of process taking place when a subject encounters new information, we constructed the series of clues for each case so that each of the nine suspects received a different combination of 0, 1, or 2 guilty clues and 0, 1, or 2 innocent clues. By looking at the rating patterns for these nine suspects, we hoped to gain some insights into the judgment processes of our subjects: Are the effects of guilty clues additive? Do innocent clues add up in the same manner as guilty clues?

Method

Subjects. Subjects were 80 male and female Northwestern undergraduate students enrolled in a general psychology course. Subjects were run in groups of 2 to 10 in single sessions lasting approximately 2 hours.

Materials. Training materials were identical to those used in Experiment 2.

The two plot scenarios from Experiments 1 and 2, "The Murdered Banker" and "The Poisoned Philanthropist," were expanded to include nine suspects. Clue sets for each story now included 23 clues; 3 of these (Clues 19, 21, and 23) were eliminators and 2 (20 and 22) were neutral. A 3 × 3 factorial design was then used to allocate the remaining 18 clues in such a way that all possible combinations of 0, 1, or 2 guilty clues and 0, 1, or 2 innocent clues were represented by the suspects.

Two separate versions of each case were constructed. In both, three suspects were eliminated: a high-probability suspect (2 guilty clues, 0 innocent clues); a medium-probability suspect (1 guilty clue, 0 innocent clues); and a low-probability suspect (0 guilty clues, 0 innocent clues). The two versions differed in the order of these three eliminations. Version 1 eliminated the three suspects in the order low, medium, high probability; Version 2 eliminated them in the order high, medium, low probability.

Design and procedure. Each subject attempted to solve both cases. The 80 subjects were divided into eight groups, differing in order of case presentation, version of the first case presented, and version of the second case presented, to counterbalance case materials and presentation order.

Procedures were essentially the same as in the first two experiments (without think alouds). To decrease the difficulty of remembering the names and identities of nine suspects, the experimenter gave each subject a list of all suspects and brief descriptions of each to refer to whenever they desired while making their ratings. Subjects were limited to 60 s for each set of nine ratings. This was ample time for all subjects to complete the task.

Results

Preelimination ratings. For both cases, final ratings (just prior to the first elimination) were averaged for each of the nine suspects. As seen in Figure 5, ratings tend to increase as the number of guilty clues increases, and ratings decrease as the number of innocent clues increases. There are, however, exceptions to this general rule. It appears that variation in the strength of individual clues is responsible for the irregularities. When we average the amount of change in the rating of the target suspect for each individual clue, obtaining a rating of clue strength, we see that an innocent clue referring to Meredith Shannon, for example, causes a much larger decrease in ratings (−.28) than do most other innocent clues (which average −.12) in "The Murdered Banker." This exceptionally strong clue is responsible for the unusually low mean rating associated with the 2-guilty-clues/1-innocent-clue value in Figure 5.

Individual subjects' ratings after each clue were again categorized in terms of adjustment to the target suspect and average adjustment to nontarget suspects. The percentage of responses qualitatively conforming to the laws of probability (i.e., involving adjustments to the nontarget suspects in a direction opposite that of the adjustment to the target) was

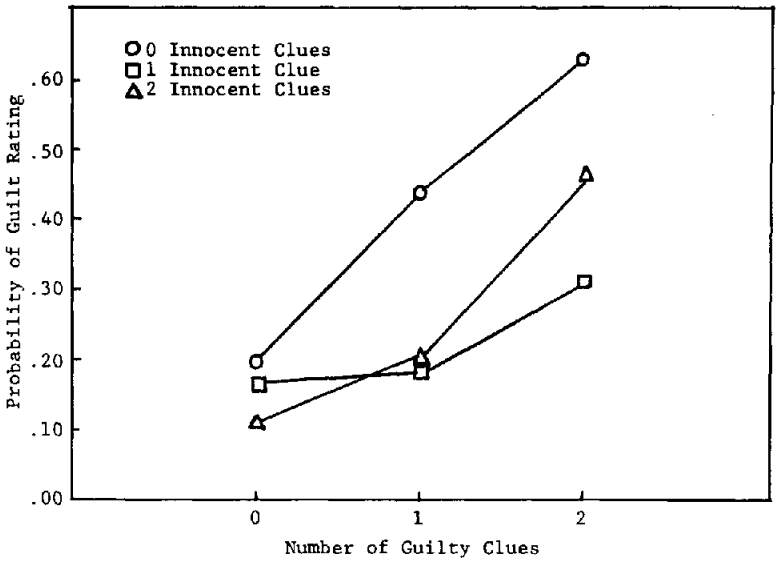

Figure 5. Mean probability of guilt ratings immediately prior to the first elimination, Experiment 3, averaged across both cases.

again consistently low: 20% for neutral clues, 23% for innocent clues, and 34% for guilty clues. Guilty and innocent clues for both cases showed strong tendencies to elicit no adjustments to nontarget suspects. It was difficult to obtain a truly neutral clue; subjects made some sort of adjustment after 80% of all neutral clues. Zero ratings and resuscitations were again common.

Postelimination ratings. Strategies used by individual subjects after an elimination were categorized as per Experiments 1 and 2. The number of uses of a particular strategy was counted and tabled as a function of (a) order of elimination and (b) probability of the eliminated suspect. Subjects' postelimination adjustment strategies (n = 160) were more predominantly flat (no adjustments to any ratings of remaining suspects) on the first and second eliminations (45% and 43%, respectively) than on the third (33%). This difference was statistically significant, $\chi^2(7, N = 480) = 23.53$, $p < .01$; however, this difference was less pronounced than that found in Experiment 2. Similar results were found with regard to the probability of the eliminated suspect; strategies (n = 160) were more often flat following elimination of a low-probability suspect (47%) than a medium-probability suspect (43%) and more often flat for the medium-probability eliminated suspects than for high-probability eliminated suspects (31%), $\chi^2(7, N = 480) = 37.49$, $p < .001$. Unlike the results of Experiment 2, however, flat was still the preferred strategy for even the high-probability eliminated suspects.

An effect of elimination sequence was also shown. Subjects' adjustments (n = 240) under the low-to-high elimination order followed the flat strategy more often (44%) than those of subjects receiving the high-to-low elimination order (37%), although this difference was only marginally statistically significant, $\chi^2(7, N = 480) = 13.59$, $p < .10$. When the sequence-of-elimination tabulation was broken down by probability of eliminated suspect, it was seen that the probability-of-eliminated-suspect effect was strong for subjects receiving Version 1, in which the low-probability suspect was eliminated first, $\chi^2(7, N = 240) = 41.67$, $p < .001$, but not statistically significant for subjects receiving Version 2, in which the high-probability elimination occurred first.

Discussion

Final ratings for the nine suspects in each case indicate that, for the most part, guilty

clues cause subjects to increase their probability ratings and innocent clues cause subjects to decrease their ratings. The patterns of responses indicate a fairly simple additive process; certain deviations, however, indicate that pieces of evidence differ in the impact they have on a judgment (see also Anderson, 1959). Whether these differences reflect differential weights of the clues (*weight* referring to the relative importance assigned to a general attribute, e.g., motive) or to their specific implications (their values on the scale from innocence to guilt) cannot be determined from our data. (For a discussion of the difference between weight and implication, see Shanteau, 1980.)

A curious tendency may be seen in subjects' responses to innocent clues. Although the presentation of one innocent clue causes a noticeable decrease in probability ratings, further innocent clues have a negligible effect. Whether this represents a general trend in the additivity of exonerating evidence or merely variations in the specific strengths of the particular clues used here cannot be determined from this study.

As in Experiment 2, subjects were more likely to use the flat strategy on early eliminations than on later eliminations. This difference, however, was much less pronounced than that found in Experiment 2. This discrepancy between the two experiments may provide some indirect evidence regarding the cause of the order-of-elimination effect as discussed previously. The major difference between the case materials used in the two experiments is their number of suspects. If the order-of-elimination effect were caused by subjects learning something over the course of three eliminations, then we would expect that same learning to occur whether the initial set of hypotheses numbered five or nine. We would not expect the effect to weaken with a larger set of suspects. Thus, this experiment provides support for the theory that the critical factor is the number of alternatives remaining at the time of an elimination.

Preelimination probabilities of the eliminated suspects were again shown to influence subjects' strategies; flat strategies were more likely when the eliminated suspect was of low probability. This difference, however, was observed only when the low-probability suspect was eliminated before the medium- and high-probability suspects. When a high-probability suspect was eliminated first, there was little difference in the proportions of flat responses among high-, medium-, and low-probability eliminated suspects. It would seem that the order-of-elimination effects and probability-of-eliminated-suspect effects can cancel each other out when in opposition; the tendency of a high-probability suspect to elicit nonflat responses is countered by the tendency of early eliminations to elicit flat responses. This suggests that the critical factor in predicting whether any adjustment will be made after an elimination is not simply the probability of the eliminated suspect or simply the number of hypotheses remaining but rather a combination of these two variables: the number of plausible hypotheses remaining after an elimination.

General Discussion

The most interesting phenomenon that we have uncovered in this research is the tendency of the adjustment process for one hypothesis to be insensitive to current evidence regarding competing hypotheses. We may say that the subject, in evaluating multiple competing hypotheses regarding the same question, treats each of the hypotheses as if it were independent of all the others. When evaluating the relative strengths of competing hypotheses, subjects seem like jurors trying several defendants charged with the same crime: Each defendant is given a separate trial in a separate courtroom, and the juror hears evidence against all defendants but is unable to use evidence gathered in one courtroom to aid a decision in any other court. Evidence favoring one hypothesis is treated as though it has no bearing on alternative hypotheses: When one hypothesis is increased in probability, others do not decrease; when one hypothesis is eliminated, others do not increase in probability. There are, however, factors that may affect the tendency to treat hypotheses as independent. Particularly after the elimination of a hypothesis, the number of plausible hypotheses remaining seems to be an important determinant of the type of

adjustments that will be made in a given situation.

Such independence is not entirely unknown in judgment literature. Teigen (1983) showed that subjects' estimates of the probability of a given hypothesis were unaffected by changes in the number of competing hypotheses with which it was being compared. This kind of independence of competing hypotheses, in Teigen's data as well as our own, is inconsistent with more than the normative model of Bayes's theorem. Most current descriptive models of hypothesis testing behavior (e.g., Anderson, 1981; Edwards, 1968; Einhorn & Hogarth, 1983; Schum & Martin, 1980) assume complementarity in probability adjustment, usually under experimental conditions where only two hypotheses are under consideration. Future models and revisions of current models will have to account for subjects' tendencies toward noncomplementary opinion revision in the multiple-hypothesis situation.

Past research has often shown that human opinion revision in a hypothesis-testing task does not necessarily follow mathematical norms; human judgment has been shown to be conservative with respect to the optimal standards of Bayes's theorem and subject to a number of biases that lead to deviations from the laws of probability theory (see Fischhoff & Beyth-Marom, 1983, for a summary review). Our results lend further support to the idea that human subjects' strengths of belief in hypotheses, though they may superficially resemble concepts of probability, actually operate according to rules quite different from the laws of mathematical probability.

Inadequate opinion revision in comparison with Bayesian theory is seen in our studies as it has been in previous research, but the discrepancy with Bayesian calculations is only one facet of the nonprobabilistic behavior shown by our subjects. Mathematical probabilities are said to exist on an "absolute" scale (a ratio scale with an endpoint), but our subjects' notion of probability seems to have properties of an ordinal scale: There is no true zero point, and these probabilities seem to have meaning only relative to each other rather than relative to some external standard unit of measurement. Sets of hypotheses that are explicitly described as exhaustive and mutually exclusive are nevertheless treated as if they are not, and the sum of probability ratings for such sets seems to have no limit. Further, when a zero probability is assigned to a hypothesis, it does not seem to mean to a subject what it would mean to a mathematician. It is instead a temporary measure, an admission that "I don't think this is possible given the information I have, but if later information makes it seem possible again, I'll reconsider it." One might say that the subject implicitly understands Keynes's distinction between the implicational force of a piece of evidence (relative to other evidence) and the weight (absolute amount) of the information currently available (see Cohen, 1977). The subject may believe that the probability of guilt, based on the implicational force of the evidence at hand, is zero, but the subject also understands that there is still more to learn and that future evidence may yet point toward guilt.

Leaving normative considerations aside for a time, it becomes important to ask not "What should subjects be doing in this task?" but "What are subjects doing?" We are not yet ready to propose a formal model of human opinion revision, but our subjects' verbalizations and ratings yield some basic insights into the process. This process may best be characterized as one of anchoring and adjustment. The subject forms an initial opinion about the likelihood of a hypothesis and then adjusts that opinion as new evidence is received (see Lopes, 1982). Adjustment processes for competing hypotheses are assumed to be independent rather than complementary. The modal strategy for adjustment can be crudely described by a simple adding rule: Positive evidence (e.g., our guilty clues) regarding a hypothesis adds a certain amount to that hypothesis's probability, further positive evidence adds a certain amount more, and so on. Negative evidence, however, may add up somewhat differently. Our data suggest that the first piece of negative evidence tends to subtract from a hypothesis's probability, but further negative evidence has little impact on the subject's opinion. It is also important to note that evidence, both positive and negative, can vary greatly in its impact on a judgment. Any given piece of evidence may be weighted heavily or only slightly,

depending on the subject's judgment of its worth.

Results of our experiments have shown that untutored subjects behave in a nonprobabilistic fashion when evaluating multiple competing hypotheses. Nonprobabilistic thinking is dramatically evident in subjects' tendencies to segregate their thoughts about the hypotheses under consideration in a manner that is objectively remarkable: Their habit is to focus on each alternative separately as if each were completely independent of all others. To categorize this behavior as irrational, however, may be unwarranted. It is hard work to evaluate competing hypotheses, and our subjects' responses may be the bottom line in a cognitive accounting as they perform a difficult task in an efficient manner given limited mental resources.

References

Anderson, N. H. (1959). Test of a model for opinion change. *Journal of Abnormal and Social Psychology, 59,* 371–381.

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

Cohen, L. J. (1977). *The probable and the provable.* Oxford, England: Oxford University Press.

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

Einhorn, H. J., & Hogarth, R. M. (1983). *Ambiguity and uncertainty in probabilistic inference.* Unpublished manuscript, University of Chicago, Center for Decision Research.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90,* 239–260.

Fischhoff, B., & Lichtenstein, S. (1978). Don't attribute this to Reverend Bayes. *Psychological Bulletin, 85,* 239–243.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1983). *Improving inductive inference through statistical training.* Unpublished manuscript, University of Michigan.

Levine, M. (1970). Human discrimination learning: The subset sampling assumption. *Psychological Bulletin, 74,* 397–404.

Lopes, L. L. (1982). *Toward a procedural theory of judgment.* Unpublished manuscript, University of Wisconsin (Madison).

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.

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

Phillips, L. D., Edwards, W., & Hays, W. L. (1966). Conservatism in complex probabilistic inference. *IEEE Transactions in Human Factors in Electronics, HFE-7,* 7–18.

Schum, D. A., & Martin, A. W. (1980). *Probabilistic opinion revision on the basis of evidence at trial: A Baconian or a Pascalian process?* (Report No. 80-02). Houston, TX: Rice University.

Shanteau, J. (1980). *The concept of weight in judgment and decision making: A review and some unifying proposals* (Report No. 228). Boulder, CO: University of Colorado Institute of Behavioral Science, Center for Research on Judgment and Policy.

Teigen, K. H. (1983). Studies in subjective probability III: The unimportance of alternatives. *Scandinavian Journal of Psychology, 24,* 97–105.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Wright, P. (1974). The harassed decision maker: Time pressures, distractions, and the use of evidence. *Journal of Applied Psychology, 59,* 555–561.

Received December 10, 1984
Revision received March 7, 1985 ■