Memory & Cognition
1990, 18 (3), 240-250

Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing

LORI ROBINSON VAN WALLENDAEL
University of North Carolina, Charlotte, North Carolina

and

REID HASTIE
Northwestern University, Evanston, Illinois

A well-documented phenomenon in opinion-revision literature is subjects' failure to revise probability estimates for an exhaustive set of mutually exclusive hypotheses in a complementary manner. However, prior research has not addressed the question of whether such behavior simply represents a misunderstanding of mathematical rules, or whether it is a consequence of a cognitive representation of hypotheses that is at odds with the Bayesian notion of a set relationship. Two alternatives to the Bayesian representation, a belief system (Shafer, 1976) and a system of independent hypotheses, were proposed, and three experiments were conducted to examine cognitive representations of hypothesis sets in the testing of multiple competing hypotheses. Subjects were given brief murder mysteries to solve and allowed to request various types of information about the suspects; after having received each new piece of information, subjects rated each suspect's probability of being the murderer. Presence and timing of suspect eliminations were varied in the first two experiments; the final experiment involved the varying of percentages of clues that referred to more than one suspect (for example, all of the female suspects). The noncomplementarity of opinion revisions remained a strong phenomenon in all conditions. Information-search data refuted the idea that subjects represented hypotheses as a Bayesian set; further study of the independent hypotheses theory and Shaferian belief functions as descriptive models is encouraged.

The literature on human judgment gives us numerous examples of human beings' failure to correctly apply statistical principles to tasks of probability estimation and opinion revision. Of particular interest are studies that address the issue of the relationship between competing hypotheses that are mutually exclusive and exhaustive. Teigen (1983), for example, has shown that subjects consistently give probability estimates that add up to more than 1.00, thus violating the "fundamental convention" of probability theory. Robinson and Hastie (1985) have presented data that challenge subjects' comprehension of a related aspect of probability theory: the assumption that, as long as a set of hypotheses is mutually exclusive and exhaustive, the likelihoods of these hypotheses should change in complementary fashion. In a series of experiments involving murder mysteries, Robinson and Hastie (1985) found that whenever a clue was evaluated, subjects tended to adjust the probability of only one target suspect, leaving the probabilities of competing suspects unchanged. Even when a suspect was completely eliminated from consideration, probabilities of the remaining suspects often were not adjusted at all. Van Wallendael (1989) found similar noncomplementary opinion revision when new suspects were added to a set already under consideration.

It is possible that noncomplementary opinion revision is merely one more "cognitive illusion" (von Winterfeldt & Edwards, 1986), one more example of naive subjects' failure to grasp the rules of probability. But we believe that the above findings apply to much more than simple probability estimation. The failure of subjects to sum their probabilities to 1.00 may indicate more than a lack of knowledge of the fundamental convention; it may indicate that subjects represent competing hypotheses not as a set, but as independent cognitive entities. As Teigen (1983) notes, "According to the classical as well as frequentistic conceptions of probability, there is a fixed probability total, 1.00 (or 100%), which has to be *distributed* over the alternatives. However, most subjects ... seem to have adopted a *non-distributional* conception of probability" (p. 104). Hypotheses (alternatives), instead of being viewed as a set sharing a fixed pool of likelihood, are mentally represented as individual possibilities, each of

This research was supported in part by National Science Foundation Grant SES-8208132 to Reid Hastie. Experiments 2 and 3 formed part of Lori R. Van Wallendael's PhD dissertation, submitted to Northwestern University. The authors are grateful to Geoffrey Fong and Peter Frey for useful advice on the research, and to Donald Mitchell and David Simkin for introducing them to Shafer's theory of belief systems. Appreciation is also extended to the participants of the Second Invitational Conference on Judgment and Decision Making, held in Nags Head, NC, in May 1988; their reactions to the first presentation of this data led to much of the theoretical work reported here. Requests for reprints should be sent to Lori R. Van Wallendael, Department of Psychology, University of North Carolina at Charlotte, Charlotte, NC 28223.

which may increase or decrease in likelihood independently of other possibilities. We shall refer to this view of hypothesis representation as the *independence hypothesis*.

The outward manifestation of noncomplementary probability revision, however, does not by itself prove that the internal representations of hypotheses are independent. To clarify this point, let us look at two formal theoretical approaches to hypothesis testing: Bayes's Theorem, and the theory of belief functions espoused by Shafer (1976). According to Bayesian probability, mutually exclusive and exhaustive hypotheses constitute a set, the members of which share a pool of probability equal to 1.00. Any piece of evidence that affects the probability of one hypothesis must necessarily affect all others in the set as well. Thus, given two equally probable hypotheses A and B, if Hypothesis A increases in likelihood to .70, the likelihood of Hypothesis B must drop to .30; the probabilities are revised in a complementary fashion because of the intrinsic relationship between A and B.

The theory of belief functions (Shafer, 1976) is also grounded in complementarity on a representational level. Beliefs exist regarding single hypotheses and also sets and subsets of hypotheses. These sets and subsets may be described in terms of a hierarchical system (see Figure 1). For example, where four hypotheses compete to explain a situation, the subject may initially invest all of his or her belief in the set {A, B, C, or D}, without assigning any of that belief to any one individual hypothesis. As evidence is acquired, belief may accumulate for {A}, for {B}, for {B or D}, and so forth. Progress toward a solution is made as belief is pushed toward lower levels of the hierarchy (individual hypotheses). The representation of hypotheses as sets and subsets is critical to Shafer's theory; indeed, evidence *against* {A} is represented within the belief system as evidence favoring {B, C, or D}. However, the outward manifestation of the system, in terms of probability ("strength of belief") ratings, will often be noncomplementary. If a piece of evidence implies that {A} is false, then strength of belief in {A} will decrease; however, since the subject assigns the complementary belief to a subset, {B, C, or D}, and not to an individual hypothesis, then the subject asked to give probability ratings for {B}, {C}, and {D} may well not change those ratings from what they were prior to the evidence.

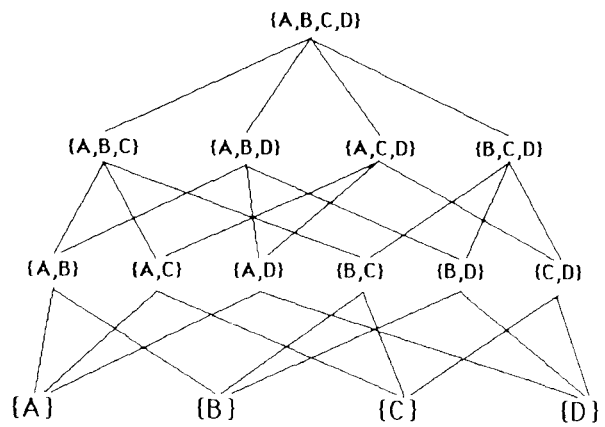

Figure 1. Hierarchy of hypothesis sets and subsets within a belief system, modeled after Shafer (1976).

Some precedent exists for proposing Shafer's system as a potential descriptive model of hypothesis testing. In particular, the theory has attracted attention in the fields of artificial intelligence and automated decision support. Mitchell and associates (Mitchell, 1987; Mitchell, Harp, & Simkin, 1987) have shown that a computer model implementing a Shaferian belief system behaves similarly to human subjects in solving the Robinson and Hastie (1985) mystery task. They have also demonstrated that subjects can interact effectively with an automated decision system using the Shafer approach. Mitchell and others argue that user acceptance of automated support depends on how natural the system's behavior appears to the human user; while Bayesian systems have thus failed to gain acceptance in many areas, the Shafer model is touted as being more promising in its similarity to human probability updating processes.

Prior research on noncomplementarity (Robinson & Hastie, 1985; Van Wallendael, 1989) has failed to support the Bayesian view of hypothesis testing as a descriptive theory of human behavior and human cognition. Several alternatives, however, have not been examined thoroughly, of which two are of interest here: (1) the Shaferian view, that the internal representation of hypotheses is basically distributive and complementary, but that subjects' overt probability estimations do not reflect this; and (2) the independence hypothesis. A test of these two theories requires a new dependent variable, since probability estimates alone do not distinguish between the two explanations. Such a dependent variable can be found in the literature on information search in choice-under-certainty tasks.

Typical experiments in information acquisition (e.g., Payne, 1976) require subjects to choose among several alternatives that vary along a certain number of dimensions; for example, apartments varying in terms of rent, noise level, size, proximity to one's workplace, and so forth. Subjects' search strategies are categorized in terms of the percentage of available information used and the order in which information is requested. It is easy to see how the principles of these studies might be applied in hypothesis-testing research. Both choices under certainty and the testing of multiple hypotheses involve combining information from various sources about several alternatives, in an effort to discover the "best" or most likely alternative of the set. Alternative hypotheses may be associated with different types of evidence (parallel to the "dimensions" variable in choice under certainty). For example, in a murder mystery, there are different kinds of information one might request about any given suspect: financial motives, opportunity, personal characteristics and background, and so forth. If we allow subjects to request such information as they test their hypotheses, we might expect their searches to fall into patterns similar

to those isolated in choice-under-certainty studies. More importantly, the amount of information requested may provide clues to the subject's internal representation of the hypothesis-testing situation.

Let us suppose that a subject is attempting to solve a murder mystery involving four suspects (hypotheses). At a certain point in the information-gathering process, a clue is revealed that eliminates Suspect A from further consideration. The elimination of a hypothesis has been used in past research as a situation in which the complementarity of a set of hypotheses is particularly relevant; however, subjects in earlier studies (Robinson & Hastie, 1985) often ignored this complementarity, making no adjustments to the probability ratings of remaining hypotheses after a non-zero-probability hypothesis was eliminated. But how will the elimination of a hypothesis affect the subject's need for information about remaining hypotheses?

If the subject has a Bayesian representation of the problem (see Figure 2A), then the elimination of A must result in increases of likelihood for B, C, and D that are proportional to their prior probabilities. Assuming that the subject will reach a decision when one hypothesis reaches some criterion probability, the elimination of A has pushed alternative hypotheses (particularly Hypothesis B) closer to criterion; thus the subject will probably require less additional information to solve the problem that he or she would have if A had not been eliminated.

If the subject has a Shaferian representation of the problem (Figure 2B), the changes in the underlying belief system will be somewhat more complicated. Elimination of A drops belief in A to zero; belief previously

A. Bayesian Representation of Probability Changes

B. Shaferian Representation of Belief Changes

C. Independence Hypothesis Representation of Likelihood Changes

Figure 2. Three models of cognitive representations for hypothesis testing.

allocated to A is now allocated to the set {B,C,D}. Also, belief previously allocated to subsets that included A will now be allocated to narrower subsets; belief in {A, B, C, or D} is now added to belief in {B, C, or D}. (For a detailed analysis of how belief systems change as evidence in a murder mystery is received, see Mitchell, 1987.) Thus, beliefs are accumulated in lower portions of the hierarchy, uncertainty is reduced, progress is made toward a solution, and the subject's need for further information should again be reduced.

Suppose, however, that our independence hypothesis describes the subject's representation of the hypothesis set (Figure 2C). No longer do hypotheses or subsets of hypotheses share one pool of belief or likelihood; instead, each hypothesis has an independent likelihood, represented here by the proportion of "pro" and "con" feelings regarding the hypothesis. Elimination of Hypothesis A in this case has no effect on beliefs regarding B, C, or D. The subject is no closer to criterion certainty, and the elimination of A has had no effect on the need for further information regarding B, C, and D.

The first two studies presented here were conducted to examine the effects of hypothesis elimination on the use of information in testing multiple competing hypotheses. The third study was designed to examine information usage and noncomplementary probability revision when evidence is set up to encourage the subject to think in terms of sets and subsets of hypotheses. Of secondary interest in all three studies were the patterns of information search in such a hypothesis-testing situation, their relationship with patterns discovered in choice-under-certainty experiments, and their consistency with Shaferian belief systems and with the independence hypothesis.

EXPERIMENT 1

This experiment was designed to explore the effect of hypothesis elimination on the need for information regarding competing hypotheses. If the subject has a Bayesian or Shaferian representation of the problem, then subjects who receive a clue eliminating one hypothesis from consideration should require less information about competing hypotheses than subjects who do not receive such an eliminator clue. However, if the representation is one of independent hypotheses, then the eliminator clue should have no effect on the amount of information requested.

A second variable explored in the experiment was the number of hypotheses (suspects) available. Prior research has indicated that this variable plays a substantial role in probability adjustment, and that the particular situation of having only two hypotheses available may be a "special case" that promotes more complementary opinion revision (Teigen, 1983; Robinson & Hastie, 1985; Van Wallendael, 1989). We chose to compare cases with four and three initial suspects, reasoning that an elimination might have a larger effect in the three-suspect case (where the elimination narrows the suspect set down to only two).

Method

Subjects. The subjects were 40 male and female undergraduate students at the University of North Carolina at Charlotte, who participated in the experiment in order to fulfill a course requirement. They were run individually, in single sessions lasting approximately 20 min.

Materials. A mystery story, "The Murdered Banker" (as used in Robinson & Hastie, 1985), was used in this experiment. The story was composed of two parts: a brief (330-word) plot scenario, which set the scene and introduced the victim and suspects, and a set of clues. The clues were of several types, providing information that might be neutral or might imply a particular suspect's guilt or innocence. Each clue could be categorized as referring to a particular target suspect and as belonging to one of five general classes of information about that suspect. For example, a "possible motives" clue regarding the victim's wife was, "An agent from the Universal Insurance Company told the police that Kitty Ostermann was the beneficiary of a $500,000 life insurance policy, to be paid on her husband's death under any circumstances." The clues were constructed so that no single clue or set of clues would logically prove one suspect to be the killer. However, taken as a whole, the clues pointed strongly toward the guilt of a particular suspect, whom most pilot subjects agreed on as the guilty party. In addition to these clues, an eliminator clue was constructed for half of the experimental conditions; this eliminator clue gave one suspect an airtight alibi and thus eliminated him or her from further consideration as a suspect in the case.

Design and Procedure. The subjects were told that they were participating in a study of judgment processes, and that their task would be to attempt to solve a brief murder mystery. Since the task would involve rating suspects' probabilities of guilt at various points in time, the concept of probability was briefly explained to each subject, and the subjects were given some sample probabilities to estimate (for example, the probability of a toss of heads on a fair coin, or the probability of rain on a given day). The subjects were instructed that in order to help them make their decisions about the probabilities of guilt for the suspects in the case, they would be given the opportunity to ask for various kinds of information about the suspects. It was stressed that although 10-20 clues would be available for each case, the subjects should attempt to solve the mystery using as little information as was needed to be "reasonably certain" about the guilty party. Equal emphasis was placed on correct identification of the killer and efficiency of information use. The subjects could choose to stop receiving clues and declare a solution whenever they wished. Finally, they were informed that one and only one of the suspects would be guilty of the crime. When this introduction to the task was completed, the subject was seated at the console of an Apple Macintosh microcomputer and told to begin when ready.

A short plot scenario for the mystery was then presented on the computer screen. When the subjects finished reading the story, they were asked to make a series of preliminary probability-of-guilt ratings for each suspect in the case, using only the information contained in the plot scenario; the subjects were instructed to rate each suspect's probability on a 0-100 scale, and to type the correct rating next to each suspect's name as it appeared on the screen. Ratings for all suspects were simultaneously visible on the screen, to minimize reliance on memory. Also, after all suspects' ratings were entered, the subjects were given the opportunity to change any or all ratings for the clue. Thus, subjects who might have wished to sum their ratings to 100 should have had ample opportunity to do so. After the initial set of ratings was made, a matrix of available clues was presented. The margins of this matrix listed the suspects in the case and five different types of information that might be requested of each suspect (see Figure 3). The subjects typed the number of the clue they wished to see next, and the requested informa-

Information Type	Suspect: Rita Frawson	Wellington Blakely	Kitty Ostermann	Vincent Carmeo
Relations With Victim	1	2	3	4
Possible Motives	5	6	7	8
Right- or Left-Handed	9	10	11	12
Whereabouts at Time of Crime	13	14	15	16
Testimony of Character Witnesses	17	18	19	20

Figure 3. Clue search matrix for "The Murdered Banker." Subjects read the plot scenario, then select clues from the matrix one at a time. The selected clue appears on the screen, after which subjects are asked to rate the probability of guilt for each suspect. The matrix reappears and subjects select the next clue.

tion then appeared on the screen. After each clue was read, the subjects were asked to again rate each suspect's probability of guilt. They were then asked if they wished to receive a new clue or to solve the mystery; if a new clue was requested, the matrix appeared again on the screen, but if the subjects chose to solve the case, they were instructed to type in the name of the guilty party. Feedback regarding the correct solution was then given to each subject.

Two independent variables were crossed in a 2×2 factorial design. The first variable was the number of suspects given at the start of the case; half of the subjects received stories involving three suspects, and half received four-suspect stories. The second variable was the presence or absence of an eliminator clue. Elimination of a suspect was accomplished by means of a "flash bulletin," which was presented (unrequested) to half of the subjects just after preliminary probability ratings were made and before any information was requested. After the elimination clue, the subjects were asked to make an additional set of probability ratings for the remaining suspects, and then they were allowed to begin with their search for information regarding the remaining suspects. The computer recorded onto a data file each subject's number of clues requested per case, the order of clues searched, and the probability ratings for each suspect after each clue.

Results

Amount of information requested. For the elimination conditions, the number of clues requested was counted for each subject and divided by the number of remaining suspects in the case to arrive at a "mean clue requests per suspect" figure. For the no-elimination conditions, clue requests counted were restricted to the suspects who were not eliminated in the elimination condition. Thus, if Rita, Kitty, and Hubert were the three suspects in the no-elimination condition, and if Hubert was eliminated in the elimination condition, then the mean clue requests per suspect for each condition reflected only requests for information regarding Rita and Kitty. A comparison of the average number of clues requested per suspect by subjects in the two elimination conditions revealed no significant differences; means were 2.37 (SD = 1.11) for the elimination condition and 1.99 (SD = 0.89) for the no-elimination condition. There was also no significant effect of number of suspects, and no interaction effect.

The total number of clues requested before declaring a solution was also examined for each condition. Number of suspects played a significant role here, with an average of 4.85 (SD = 2.50) clues requested for three-suspect groups and 7.20 (SD = 2.98) for four-suspect groups [$F(1,36) = 7.12$, $p < .011$]. Again, there was no significant effect of an elimination.

Adjustments after nonelimination clues. Out of 241 clues received by all subjects, 52 clues elicited no adjustments to any suspect's rating. Of the 189 clues on which adjustments were made, 35% involved adjustments to the probability of only the target suspect, and none of the competing suspects; 5% involved increases to the ratings of two or more suspects with no complementary decreases to alternatives; 8% involved decreases to the ratings of two or more suspects with no complementary increases; and 52% involved combinations of increases and decreases to suspects' ratings (although rarely in Bayesian proportions). Thus, almost 50% of probability adjustments were noncomplementary in qualitative terms, and almost all were non-Bayesian in quantitative terms. There were no significant differences in percentage of noncomplementary revisions due to the number of suspects or to the presence/absence of eliminator clues.

Revisions after eliminator clues were not included in the above analysis, for two reasons: (1) Such clues were not experienced by half of the subjects; and (2) since eliminators were nonrequested clues, they might be expected to differ from requested clues in ways that would destroy their comparability. The reader who is interested in the effect of eliminations on probability ratings is referred to our earlier paper (Robinson & Hastie, 1985).

Correlations between the amount of adjustment to target suspects and the amount of adjustment to nontarget suspects were also calculated. Subjects following the laws of Bayesian probability theory should show a perfect −1.00 correlation between the adjustment to the target suspect and the summed adjustments to nontarget suspects for each clue, with a slope of −1.00 and an intercept of 0.00. Our subjects fell short of this optimal performance in all conditions; mean correlations and slopes were −.196 and −.035 for the four-suspect no-elimination condition; −.189 and −.343 for the four-suspect elimination condition; −.064 and −.093 for the three-suspect no-elimination condition; and −.415 and −.387 for the three-suspect elimination condition. Only the correlation for the three-suspect elimination condition ($t = -2.09, p < .07$) and the slopes for the four-suspect elimination condition ($t = -2.68, p < .03$) and the three-suspect elimination condition ($t = -1.98, p < .08$) tended to differ significantly from zero.

Information-search patterns. Individual clue requests were analyzed in two ways. First, clues requesting information about either the same target suspect or the same information type as the previous clue were counted and tabulated. Second, the target of each clue request was noted, and patterns of searching for information about favorite and longshot suspects were examined. The subjects showed an overall tendency to favor within-suspect search (50% of all clue requests) over within-clue-type search [28%; $\chi^2(1) = 12.41, p < .001$]. With respect to to the targets of requested clues, the subjects overwhelmingly preferred information regarding their current favorite suspect (55% of all clue requests) as opposed to longshot suspects [28%; $\chi^2(1) = 21.78, p < .001$]. "Favorite" and "longshot" are here defined relative to the individual subject; thus, Subject A's highest rated (favorite) suspect after Clue 4 may not be the same as Subject B's "favorite."

Discussion

Amount of information requested. No significant differences in information use were found between the elimination and no-elimination conditions. This again argues against a Bayesian representation of the problem. A Shaferian representation is also unlikely to result in such behavior. However, since the elimination occurred at the very beginning of information search, it is theoretically possible that belief was concentrated solely in the realm of "uncertainty" (Hypothesis {A, B, or C}), and that the restructuring of beliefs after the elimination was not significant enough to warrant a noticeable difference in information usage. Thus, although the behavior of at least half of our subjects seems consistent with the independence hypothesis, the Shaferian model cannot yet be ruled out.

Noncomplementary opinion revision. As in previous research (Robinson & Hastie, 1985), the subjects showed a marked tendency to adjust only the probability of a clue's target suspect and not of any of the competing suspects. Correlations between the amount of adjustment to target suspects and nontarget suspects were generally negative, but substantially less so than the Bayesian prediction of −1.00, and the slopes of the regression lines indicated that the subjects were making much smaller adjustments than Bayesian probability theory would dictate. The major exception to the rule is the three-suspect elimination condition. In this condition, an initial set of three suspects is narrowed down to two by the elimination. Past research (Teigen, 1983; Robinson & Hastie, 1985; Van Wallendael, 1989) has suggested that the two-hypothesis case may be the only situation in which many subjects adopt a complementary, distributive approach to probabilities; when the set is narrowed down to this extent, a Bayesian or Shaferian representation of the hypothesis set may be used.

Information-search patterns. The popularity of within-suspect patterns of clue choice lends some support to the notion that subjects are considering the hypotheses as independent entities. Likewise, the focus on favorite suspects is consistent with the idea of attempting to achieve a criterion probability before making a decision.

EXPERIMENT 2

The Shaferian representation of belief is likely to be influenced not only by the presence of an eliminator clue, but also by its timing. As alluded to earlier, an elimination in the early stages of information gathering may simply redistribute belief from one "uncertain" set {A,B,C,D} into a slightly smaller subset {B,C,D}; lower levels of the belief hierarchy may be relatively unaffected. But after more information has been gathered, those lower levels of the hierarchy should have some belief associated with them; to be specific, the more information has been gathered, the more likely it is that there is some nonzero amount of belief associated with the {A,B} subset that can be redistributed to {B} upon the elimination of Hypothesis A. Thus, a later elimination would be more likely to elicit complementary probability adjustments, and also to have an impact on the amount of information necessary to make a decision. However, the independence model would predict no effects of elimination timing on information use. The following experiment was designed to test the opposing predictions of the Shaferian and independence theories.

Method

Subjects. The subjects were 120 male and female undergraduate students at Northwestern University, who participated in the experiment in order to fulfill a course requirement. They were run individually, in single sessions lasting approximately 45 min.

Materials. Two mystery stories were used: "The Murdered Banker," as described in Experiment 1, and "The Poisoned Philanthropist," another case used by Robinson and Hastie (1985). Each case involved five suspects and five clue types, for a total of 25 available clues. Three versions of each case were constructed: one in which an eliminator clue was presented after 5 requested clues (early), one in which the eliminator was presented after 10 requested clues (late), and one in which the eliminator was presented after 15 requested clues (very late). Only one eliminator clue was included in each case.

Design and Procedure. The procedure was identical to that in Experiment 1. The subjects were told that they would be attempting to solve two mysteries that might involve varying numbers of suspects and clues. It was again stressed that they should attempt to solve each case using as little information as was needed to be reasonably certain of the guilty party, and they were again informed that one and only one of the suspects would be guilty in each case. An Apple III microcomputer controlled the presentation of stories and clues.

Each subject was run in two of the three elimination conditions; thus, a subject might be exposed to a late elimination in the first case and an early elimination in the second case. In all, 80 subjects were exposed to each of the three conditions, with order of condition and case presentation counterbalanced across subjects. The computer again recorded the number of clues requested per case, the order of clues searched, and the subject's probability ratings for each suspect after each clue.

Results

Amount of information requested. On the average, subjects took 12.74 clues before declaring a solution for each case. Notice that this represents fewer than the 15 clues that were needed to reach the eliminator clue in the very late elimination condition. Consequently, only 15 of 80 subjects in this condition actually received an elimination, which rendered the very late condition uninterpretable; hence only the early and late conditions are analyzed and reported below. (Very few subjects in the early and late conditions declared a solution before the eliminator clue had been received; these few were not removed from the analyses below.)

A comparison of the number of clues requested by subjects in the early and late conditions reveals no significant differences (means equal 13.22 and 12.42, with *SD*s of 4.93 and 4.16, respectively). The average number of clues taken by subjects in both conditions for the first case was 12.36 ($SD = 4.72$); the average for the second case was 13.29 ($SD = 4.39$); this difference also was not statistically significant.

Adjustments after nonelimination clues. Noncomplementary opinion revisions occurred on an average of 47% of all clues on which revisions were made ($n = 1{,}581$). There were no significant differences between the two elimination conditions in terms of percentage of noncomplementary revisions. The large number of subjects in this experiment made it possible also to investigate the time course of noncomplementary revisions. As seen in previous studies (Van Wallendael, 1989), the percentage of noncomplementary revisions (relative to all clues received) decreased significantly as more clues were obtained (see Figure 4); there was a significant negative correlation ($r = -.68$, $p < .01$) between clue ordinal position and percentage of noncomplementary revisions occurring.

Correlations between the amount of adjustment to target suspects and the amount of adjustment to nontarget suspects were also calculated. There were no significant differences between the two elimination conditions; the subjects fell short of the Bayesian norm ($r = -1.00$) in both conditions. For the early elimination condition, $r = -.30$, with a slope of $-.32$ and an intercept of -0.01; for the late condition, $r = -.28$, with a slope of $-.33$ and an intercept of -0.03.

Information-search patterns. The subjects showed equal overall tendencies toward within-suspect (30% of all clue requests) and within-clue-type (31%) choices in their searches for information. The timing of the eliminator clue did not seem to affect these search patterns. With respect to the targets of requested clues, the subjects again preferred information regarding their current favorite suspect (41% of all clue requests) as opposed to longshot suspects (19%).

Discussion

Amount of information requested. No significant differences were found between the early and late elimination conditions. As in Experiment 1, this finding ar-

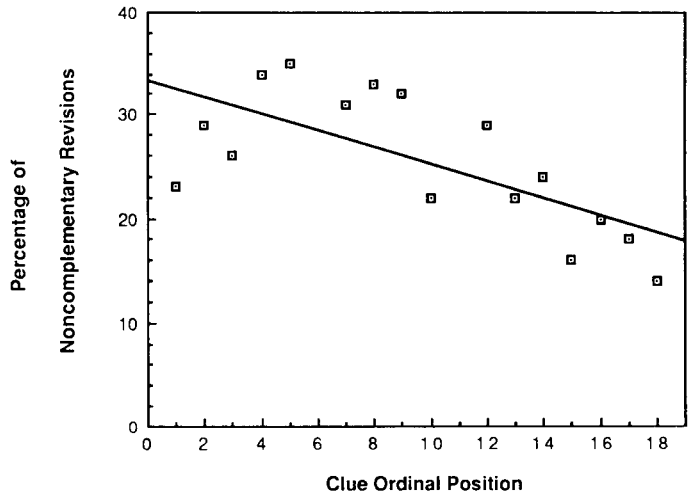

Figure 4. Percentage of noncomplementary probability revisions across all subjects for each clue received, Experiment 2.

gues against a Shaferian belief system structure as being the cognitive basis of opinion revision. In such a belief system, a late elimination should have an impact on belief in other individual hypotheses, and thus it should speed the subjects toward solution. If the subject's representation of the problem is a system of independent cognitive units, however, the elimination of one unit, whenever it occurs, will not have any impact on belief in other units.

Noncomplementary opinion revision. As in the previous experiment, subjects showed a strong tendency to adjust only the probability of a clue's target suspect and not of any of the competing suspects, regardless of elimination condition. Note, however, that the percentage of noncomplementary revisions decreases as more information is received. This suggests one of two explanations: On the one hand, the finding is consistent with a Shaferian belief system; the more information is accumulated, the more likely it is that belief is invested in subsets of two hypotheses (e.g., {A or B}). Evidence against B will have a direct impact on belief in A when there is already some belief invested in {A or B}. On the other hand, it may be that the subject has an independent representation of hypotheses, but is gradually ruling out some of the suspects as evidence accumulates; when there are only a small number of hypotheses left (e.g., two), the subject changes to a different problem representation—a more Bayesian or Shaferian representation. The idea of changing strategies as information search progresses is well accepted in other judgment domains; for example, in the choice-under-certainty literature, Payne (1976) showed that some subjects switch from a noncompensatory to a compensatory decision rule after the information set has been narrowed to a manageable size.

Information-search patterns. Unlike the subjects in Experiment 1, these subjects did not seem to favor within-suspect clue searches over other clue-search patterns. This may simply reflect a difference between the university populations from which the two subject samples were drawn. Both groups of subjects did, however, prefer information about the current favorite suspect, consistent with the proposed goal of satisfying a subjective confidence criterion before making a decision.

EXPERIMENT 3

The results of Experiments 1 and 2 suggest that subjects' need for information is insensitive to eliminations of hypotheses. This would seem to support the view that subjects represent the given hypotheses independently from one another. However, there are alternative explanations for such reactions to an elimination. One potential problem involves the elimination paradigm in general, and particularly the use of information requests as a dependent variable. The prediction of the three models regarding information requests are based on the assumption that the criterion in question is an absolute probability level. It is reasonable, however, that a subject's probability of making a guess at the solution is based not only on the absolute likelihood of the favorite suspect, but also on this suspect's relative likelihood as compared with that of other suspects. For example, if Suspect A has an assigned probability of .50, and if Suspects B, C, and D each have probabilities of .17, the subject may guess that A is guilty; however, if the probabilities for B, C, and D are .49, .005, and .005, the subject may opt for further information. The effect of an elimination on the need for further clues, then, would depend on the particular suspect eliminated and the distribution of probabilities before the elimination; thus, predictions of the need for information become much more complex under all hypotheses.

This "relative-criterion" problem can be addressed by the data. For Experiment 1, we looked at the subjects' absolute probability ratings for the suspect chosen as guilty at the time of choice. These ratings varied from 50 to 100, with a mean value of 90.925 (SD = 14.17, N = 40). By finding the ratio of (chosen suspect's probability)/(sum of probabilities of the two highest rated suspects), we also looked at the relationship between the rating of the chosen suspect and the rating of its nearest competitors. These ratings were much more variable, with a range of 43 to 100 and a mean of 72.675 (SD = 20.99). It was actually fairly common for a subject to choose Suspect A at a probability of .90, even when Suspect B was rated at .80 or .85. Such findings are not consistent with the notion of a relative decision criterion. However, the findings are consistent with an absolute criterion for most subjects; 32 out of 40 made their decision after the probability rating of the chosen suspect reached .90 or higher.

Even if we concede that the subject's criterion is absolute, however, another problem remains. It is possible that, when a hypothesis is eliminated, subjects realize that the problem has become easier, but react to the simplification by changing their criteria for making a decision. They may reason that, since the problem is now made simpler, they can "afford to" request more information and be more certain of their choice before declaring a decision. The notion of such a criterion shift can be informally tested by looking at the percentage of subjects who choose the consensus "best" suspect as the guilty party; if hypothesis eliminations do cause a criterion shift, then subjects who receive an elimination might pick the "correct" answer more frequently, since they have adopted a higher decision criterion and are less likely to be fooled by a few pieces of evidence against an innocent person. Such a difference does exist in the data from Experiment 1, where 35% of the subjects in the elimination condition picked the correct choice as opposed to 25% in the no-elimination conditions; neither of these percentages, however, differs significantly from chance accuracy. A more compelling pattern is seen in Experiment 2; 70% of the early elimination subjects, 57% of the late elimination subjects, and only 26% of the no-elimination subjects chose the correct suspect (chance accuracy for all conditions being 20%). The idea of a criterion shift, then, cannot be lightly dismissed. Further research is needed, perhaps utilizing instructions and payoffs that place greater

stress on using the minimal number of clues. Meanwhile, however, it would be useful to examine subjects' behavior in an information-search situation in which the amount of material that may be requested is set by the experimenter, not by the subject; in such a situation, subjective criterion shifts would not be a problem.

One further issue needs to be addressed at this time. In the previous two studies, as well as studies reported in earlier papers (Robinson & Hastie, 1985; Van Wallendael, 1989), clues have been utilized that may have encouraged subjects to adopt a view of the hypotheses as being independent. These clues typically mention one and only one suspect by name—for example, "Kitty had recently taken out a $100,000 life insurance policy on her husband." Even if a subject's natural tendency were to adopt a Shaferian representation, such clues might push the subject toward noncomplementary revisions of probability estimates. Suppose, however, that clues referred to subsets of suspects rather than single suspects—for example, "A blonde hair was found on the victim's coat; both Kitty and Rita have blonde hair." If the representation was most like a Shaferian belief system, then such a clue would add belief to the {Kitty or Rita} hypothesis; subsequent clues implying Kitty's innocence would then be reflected in increases in Rita's probability of guilt, assuming that the subject has reasonably good memory for the prior information. In general, a case involving several such multiple-target clues should encourage greater complementarity in probability revisions if the representation is Shaferian. If the independence hypothesis is true, however, multiple-target clues will affect the probabilities of only the targets named within the clue, and they will not have any particular effect on reactions to later evidence received. Experiment 3 was conducted to test these predictions.

Method

Subjects. The subjects were 120 male and female undergraduate students at Northwestern University, who participated in the experiment in order to fulfill a course requirement. They were run individually in single sessions lasting 25–50 min.

Materials. Two mystery stories were used, as described in Experiment 2. Each case involved five suspects, and a total of 25 clues of five different types were available in each clue matrix. Two types of clues were now used: the standard single-target clues used in Experiments 1 and 2, and also multiple-target clues, which contained information referring to the suspect listed in the matrix plus at least one other suspect. For example, a subject requesting a clue about Alice Robner's knowledge of poisons might receive the following multiple-target clue: "Alice Robner has a passing knowledge of herbicides, acquired during her summers working in the Robner greenhouse. Brad Michaels [the gardener] has taught her a great deal about the most efficient weed killing methods."

Three versions of each case were constructed: (1) a control condition, in which all clues were of the standard single-target type; (2) a low-percentage condition, in which 5 (20%) of the available clues were multiple-target clues; these clues were substituted for standard clues in such a way that each primary suspect and each clue type in the matrix was represented by 1 multiple-target clue; and (3) a high-percentage condition, in which 10 (40%) of the available clues were multiple-target clues, with each primary suspect and each clue type being represented by 2 multiple-target clues. Each version of each case contained 1 eliminator clue, which was presented after the 10th clue.

Design and Procedure. The subjects were assigned to either a search group or a yoked control group. The search group was allowed to request the specific clues desired; the yoked control group received the sequences of information requested by the search group, but received them passively. This was done to provide a direct test of differences elicited by information-search procedures as opposed to the more passive procedures used in past research (Robinson & Hastie, 1985; Van Wallendael, 1989). The procedure for the search group was similar to those described in Experiments 1 and 2. The subjects were told that they would be attempting to solve two mysteries that would each involve five suspects. The procedure for choosing clues from the matrix was explained, and the subjects were told that they would be allowed to choose exactly 15 clues (no more or less) for each case before the solution would be presented. The subjects were not warned that some of the clues might refer to more than one suspect. It was stressed that one and only one of the suspects in each case would be guilty of the murder.

Each of the 60 subjects in the search group was randomly assigned to one of the three multiple-target clue conditions. The first case solved conformed to the subject's experimental condition—that is, the first case was a control case for 20 subjects, a low-percentage case for 20 subjects, and a high-percentage case for 20 subjects. The second case was always a control case. This was done in order to see if any effects of experiencing multiple-target clues would then transfer to a second case involving only single-target clues. The order of case presentations was counterbalanced; the computer recorded the order of clues searched and the subject's probability ratings for each suspect after each clue.

Within the yoked control group, each subject was randomly paired with one of the 60 search subjects. Yoked subjects were given the same instructions and treatment as search subjects, except that yoked subjects were not allowed to choose which clues they wanted to see. Yoked subjects saw no clue matrices; instead, they simply saw the plot scenario on the screen, made their preliminary probability ratings, and then were presented with the sequences of 15 clues that had been chosen by the search subjects to whom they were yoked. Probability ratings were made after every clue seen, and the eliminator clue was presented after the 10th clue, just as for the search subjects. The computer recorded each subject's probability ratings for each suspect after each clue.

Results

Adjustments after nonelimination clues. The mean number of noncomplementary revisions per subject was calculated for each experimental condition. (For multiple-target clues, a set of revisions was classified as noncomplementary if there were no adjustments made to nontarget suspects and the sum of adjustments to the targets did not equal zero.) There were no significant differences found due to the percentage of multiple-target clues (means were 4.00, 4.25, and 4.02 for the control, 20%, and 40% groups, respectively), the experimental paradigm (4.39 for the search group, 3.79 for the yoked group), or the order of case presentation (4.02 for the first case, 4.17 for the second). There were also no significant interactions found.

Correlations between the amounts of adjustment to target and nontarget suspects were also calculated. Again,

the subjects in this experiment fell short of the Bayesian correlation of −1.00; correlations for the various experimental conditions ranged from −.14 to −.45. For the first case solved, increasing the percentage of multiple-target clues had no systematic effect on the search group (r = −.24, −.32, and −.28 for the control, 20%, and 40% conditions, respectively; n = 300 for all conditions) and a detrimental effect on the yoked group (rs of −.45, −.33, and −.24, respectively). For the second (control) case, having solved a previous case with either 20% (r = −.40) or 40% (r = −.41) multiple-target clues improved correlations for the yoked group (r for the control condition being −.33), but previous experience with a case involving 20% multiple-target clues was associated with the poorest correlation (−.14) for the search group (with r = −.34 for the control condition and −.35 for the 40% condition).

Even when we look within the multiple-target clues only, noncomplementary opinion revision is prominent. Correlations between summed adjustments to targets and summed adjustments to nontargets are not significantly different from zero for multiple-target clues that make up 20% of the available information for a case (r= −.02 for the search group, .14 for the yoked group; n = 63), and they improve only slightly for multiple-target clues making up 40% of the available information (r = −.28 for the search group, −.22 for the yoked group; n = 126).

Information-search patterns. Individual clue requests were analyzed as in Experiments 1 and 2. The subjects in the control condition (no multiple-target clues) showed equal overall tendencies toward within-suspect (29%) and within-clue-type choices (29%). The subjects who were exposed to multiple-target clues, however, showed a slight preference for within-clue-type search (34% for the 20% condition, 37% for the 40% condition) over within-suspect search (27% and 26%, respectively). With respect to the targets of requested clues, the subjects again preferred information regarding the current favorite suspect (37%) over information about the longshot (19%).

Discussion

Noncomplementary opinion revision. The addition of multiple-target clues to the murder-mystery task had a minimal effect on noncomplementary opinion revision. In terms of responses to noneliminator clues, the lack of significant differences in the number of noncomplementary revisions per subject is further evidence of the strength and persistence of the noncomplementarity phenomenon. Correlations between the amounts of adjustment to target and nontarget suspects were not significantly improved by the inclusion of multiple-target clues; indeed, in one case (the 20% multiple-target-clue group within the search condition, second case solved), experience with multiple-target clues may have actually worsened performance. It is unclear why this particular group showed such a low correlation as compared with other groups; simple statistical variation may be responsible. In any case, the overall patterns of noncomplementarity again support the independence hypothesis as a plausible model for subjects' representations of the hypothesis-testing situation.

Information-search patterns. In contrast to the subjects in Experiments 1 and 2, the subjects in Experiment 3 were operating under subtly different instructions; instead of being asked to solve the case using as few clues as were necessary, the subjects in Experiment 3 were given a fixed number of clues that they would be allowed to utilize, and they had to choose neither more nor less than the fixed number of clues. Nevertheless, the subjects still consistently preferred information regarding the current favorite suspect to information regarding longshots, again lending credence to the absolute confidence criterion.

CONCLUSIONS

We have seen in these three studies that noncomplementarity is a robust phenomenon. It persists when subjects are allowed to search for their own information regarding the hypotheses under consideration; it is not an artifact of the passive clue-reception paradigm. This is an important finding, since in allowing subjects to search for information in whatever order they chose, we allowed them to make direct comparisons between suspects: Was Kitty's motive as strong as Rita's? Which suspects had opportunity to poison the victim's tea? Many subjects did make such comparisons in their searches for information, as is shown by the prevalence of within-clue-type choice patterns. However, even though subjects seemed to recognize that a series of related clues could be used to compare suspects in this manner, many still did not recognize the more subtle point that any single clue must have implications for all members of the set, not merely its target member. The inclusion of multiple-target clues also did little to improve the subjects' notions of relationships among hypotheses.

Given the robust nature of the effect, it is interesting to speculate about what leads people to treat these nonindependent hypotheses as if they were independent. It might be, of course, that the "real world" encourages such a view; most hypothesis-testing situations do not involve clear sets of mutually exclusive and exhaustive possibilities, and as such, they do not lend themselves to a Bayesian complementary analysis. Treating the hypotheses as independent is not necessarily suboptimal if one does not have an exhaustive set, and if subjects are accustomed to representing real-world hypotheses as independent entities, they may continue to use such a representation even when other representations are appropriate and more efficient. An alternative answer, however, concerns the cognitive capacity required by the Bayesian and Shaferian representations, both of which require a great deal of memory and computation. Treatment of hypotheses as independent may simply be more suited to our limited information-processing capacity.

Taken together, the three experiments reported here provide support for the independence hypothesis as a model

of cognitive hypothesis representation. We recognize that the evidence is hardly conclusive, since the independence hypothesis predicts null effects in all three of these studies. However, we feel it is unlikely that a simple lack of power is responsible for the failure to find significant differences in these three experiments. First, a large number of subjects and a variety of different manipulations were involved in these studies, yet no difference came close to signficance. Second, the fact that results of these three studies are quite consistent with past research indicates that the paradigm is not faulty. While a Shaferian belief system representation can account for some of the noncomplementary opinion revision seen in previous studies (Robinson & Hastie, 1985; Van Wallendael, 1989), it is inconsistent with the information-use data and complementarity data reported here.

However, it is important to note that the Shaferian model provides a better account than the independence model does for one particular finding: the decrease in noncomplementarity as information builds up (Figure 4). This phenomenon follows naturally from the Shafer model; the independence hypothesis, on the other hand, must posit a shift in representational strategy to explain these results. Perhaps a different mode of subject response might allow support for the Shafer model to surface; for example, if subjects were free to distribute probability any way they wished and were not confined to reporting probability estimates for individual suspects, would Shafer's superset/subset structure appear in the data?

If, on the other hand, the independence hypothesis is correct, further studies are needed to address the apparent change in representation that occurs as information accumulates. Is it a result of the subject's attempt to narrow down the set of plausible alternatives? This would be consistent with the general view of limited information-processing capacities. However, we do not know just what those limits are; does the representational change only occur when the set is narrowed to two hypotheses, or may it occur earlier for some subjects? And if such a switch occurs, what is the new representation—Bayesian, Shaferian, or something else? The fact that human beings revise probability estimates in a noncomplementary way has been amply demonstrated; it is time now to concentrate our efforts on the cognitive representations and processes that underlie our uses and misuses of probability.

REFERENCES

MITCHELL, D. H. (1987). *Automated decision support using variations on the Dempster-Shafer theory*. Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

MITCHELL, D. H., HARP, S. A., & SIMKIN, D. K. (1987). A knowledge-engineer's comparison of three evidence aggregation methods. In *Second Workshop on Uncertainty in AI* (pp. 297-304). San Mateo, CA: Morgan Kauffman.

PAYNE, J. W. (1976). Task complexity and contingent processing in decision-making. *Organizational Behavior & Human Performance*, **16**, 366-387.

ROBINSON, L. B., & HASTIE, R. (1985). Revision of opinion when a hypothesis is eliminated from consideration. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 443-456.

SHAFER, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

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

VAN WALLENDAEL, L. R. (1989). The quest for limits on noncomplementarity in opinion revision. *Organizational Behavior & Human Decision Processes*, **43**, 385-405.

VON WINTERFELDT, D., & EDWARDS, W. (1986). *Decision analysis and behavioral research*. Cambridge, England: Cambridge University Press.

(Manuscript received May 25, 1989;
revision accepted for publication September 19, 1989.)