limitations. In addition, subjects will be asked to identify the best rule among our set of proposed strategies.

Subjects for this experiment will be male and female college students, since our past research suggests that this age group should provide substantial numbers of a versus b, sum of diagonals and conditional probability judges. Sex of subject will be considered as a factor in the design in light of common findings of sex differences in math skills among adolescents and adults (e.g., Maccoby & Jacklin, 1974).

Method

Subjects

Subjects in the experiment were students in an introductory psychology class who participated in the experiment as one option in fulfillment of a course requirement. Subjects ranged in age from 18 to 32 years, with a mean age of 19.42. Sixty-two female and 54 male students participated.

Problems

Subjects judged a set of 12 covariation problems, structured so that each of four judgment rules would produce a distinctive judgment pattern on a problem set. Table 1a lists the actual problems used. The 12 problems include three problems for each of the four strategy types. One noncontingent and two contingent relationships are included for each strategy problem type.

Twelve different problem contents were developed, each of which consisted of a set of observations picturing one of two states for two potentially related everyday events. Three problems pictured bakery products which either rose or fell in association with the presence or absence of yeast, baking powder, or a "special ingredient." In three other problems, plants were pictured as healthy or sick as a possible function of the presence

Final Report

National Institute of Education
Grant NIE-G-80-0091

Children's Judgments about Covariation Between Events: A Series of Training Studies

Harriet Shaklee, Principal Investigator
Department of Psychology
University of Iowa

Table of Contents

(Appendices A, C, D, and E have also been processed as individual documents.)

Rationale

A much neglected area of research in mathematical reasoning is that of children's understanding of statistical concepts. Statistical problems, however, do stand as prime areas for application of mathematical training. In particular, statistics are necessary for identifying predictability in an environment where relationships are frequently probabilistic (x is more likely when y is present) rather than deterministic (x always occur when y is present). Problems such as these are common in identifying regularities in scientific phenomena, and in everyday contexts as well. In this respect, statistics provide a key link between basic mathematical concepts and central aspects of scientific and everyday problem solving.

As an area for application of mathematical training, research on statistical reasoning may also be informative about children's ability to apply their mathematical skills appropriately. Central to probabilistic reasoning is understanding of ratios and fractions. Since a probability is a ratio between two frequencies, probability assessment requires that a person be able to identify the relevant frequencies and calculate the ratio between them. Thus, research in statistical reasoning should prove profitable in understanding children's acquisition of basic skills as well as their ability to use those skills in applied settings.

Reasoning such as this underlies the call of several educators for development of training programs to improve children's understanding of statistical concepts (e.g., Harvey, 1975; Cambridge Conference on the Correlation of Science and Math in the Schools, 1969). Research in this area is critical for developing and testing such curricula in probability and statistics (e.g., Shepler, 1969; Kurtz & Karplus, 1979; Ojeman, Maxey, & Snider, 1965 a & b) for children in the elementary through high school years.

The focus of existing research in this area has been on children's probability judgments. Early work by Piaget and Inhelder (1975) indicated that full understanding of probability was realized by adolescence. Subsequent work by other investigators indicates that younger children evidence some preliminary concepts of probability (e.g., Fischbein, 1975; Yost, Siegel, & Andrews, 1962; Goldberg, 1966), and that training is effective in improving their judgments (Ojeman, Maxey, & Snider, 1965 a & b; Shepler, 1969; Dunlap, 1980).

A statistical judgment more common in causal reasoning builds on probability assessments of this sort. An individual investigating the relationship between potential cause x and effect y would compare the likelihood of x occurring when y is present $P(x/y)$ with the likelihood that x occurs without y $P(x/\bar{y})$. The two events are independent if these conditional probabilities are equal; nonindependence is indicated by any difference. The comparison is made to identify contingency or covariation between events. Scientific procedure and statistical analyses testify to the key role of covariation analyses in professional practice. Although not sufficient for causal inference, covariation is a necessary condition between cause and event. Many psychologists further assert that everyday causal judgment is similarly based on a covariation analysis (e.g., Michotte, 1963; Inhelder & Piaget, 1958; Kelley, 1967; Heider, 1958). That is, people search for likely

explanations of everyday events by identifying event covariates. Thus, competence at covariation judgment may determine a person's adequacy at identifying real world cause-effect relationships.

Unfortunately, research investigating people's competence at judging covariations between events has resulted in a maze of contradictory results. In the basic paradigm, subjects are presented with data instances illustrating one of two event states (e.g., presence or absence) for each of two events. The subject's task is to identify the direction and/or strength of the relationship between the events. Inhelder and Piaget (1958) and Seggie and Endersby (1972) each found accuracy to be the norm among adolescent and adult subjects identifying such relationships. Others (e.g., Niemark, 1975; Smedslund, 1963; Jenkins & Ward, 1965; Adi, Karplus, Lawson, & Pulos, 1970) have found full competence to be rare among populations comparable in age and expertise.

While the evidence indicates that covariation judgments are often erroneous, those judgments may be rule-governed nonetheless. Specifically, subjects may evaluate relationships according to a variety of rules, each of which should produce a characteristic performance pattern. Four rules are proposed as possible judgment strategies. The rules are discussed in terms of possible relationships between two events (A and B), each of which occurs in one of two states (1 and 2). Possible combinations of those event states are illustrated in Table 1.

Least sophisticated of the proposed strategies is judgment according to the frequency with which the target events cooccur (A_1B_1, cell a in Table 1), failing to consider joint event nonoccurrences (A_2B_2, contingency table cell d) in defining the relationship. A subject using this strategy would identify a positive relationship between A_1 and B_1 if cell a frequency was the largest of the contingency table cells, a negative relationship if it was the smallest (cell a strategy). This strategy is identified by Inhelder and Piaget (1958) as common among younger adolescents. Smedslund (1963) suggests that the strategy is typical among adults as well. The strategy does consider some relevant information and may result in better-than-chance performance. However, the rule is a limited one, and would be especially misleading when there is a large difference between frequencies in contingency table cells a and d.

A much improved approach would be the strategy defined by Inhelder and Piaget (1958) as characteristic of formal operational thinking. Specifically, covariation would be defined by comparing frequencies of events confirming (cells a and d) and disconfirming (cells b and c) the relationship. Thus, the rule would compare the sums of the diagonal cells in the contingency table (sum of diagonals strategy). Jenkins and Ward (1965), however, suggest that this strategy has its limits as well. Specifically, the rule is an effective index only when the two states of at least one of the variables occur equally often. Otherwise, a correlation may be indicated when, in fact, independence is the case.

Instead, Jenkins and Ward (1965) suggest that covariation is more appropriately evaluated by comparing the probability of event A_1 given event B_1 $P(A_1/B_1)$ with the probability of A_1 given that B_2 has occurred $P(A_1/B_2)$. This is equivalent to a comparison of the frequency ratio in Table 1 cells $\frac{a}{a+c}$ with that in cells $\frac{b}{b+d}$. By definition, independence is indicated by equivalence between these conditional probabilities; non-independence is indicated by any difference (conditional probability

Table 1

	B_1	B_2
A_1	a	b
A_2	c	d

strategy). This is the most sophisticated of our proposed strategies and should result in accurate judgment of any contingency problem.

Thus, four alternative strategies were proposed to account for subjects' judgment patterns. According to the analysis a subject's error rate should depend on the particular correlation problem he or she is judging. Problems could be identified which would be accurately judged by all four strategies. Alternatively, error rates may be high on problems solved only by the more general strategies.

This analysis suggests a powerful tool for identifying strategies actually used in covariation judgment. Since different rules produce different judgments, covariation problems might be identified which would differentiate between those rules. In fact, careful structuring of a problem set would allow us to identify the specific strategy a subject is using.

A set of such problems is illustrated in Table 2a. Problems are structured hierarchically such that cell a problems are correctly solved by all strategies; a versus b problems are correctly solved by a versus b, sum of diagonals and conditional probability strategies. Sum of diagonal problems will be accurately judged by sum of diagonal and conditional probability strategies. Conditional probability problems would be correctly solved by the conditional probability strategy alone. Solution accuracy is indexed by the direction of the judged relationship (i.e., A_1 more likely given B_1, B_2, or no difference). A subject's solution pattern on the set of problems indicates the strategy used. Problems on the first row of Table 2a illustrate judgments predicted by each of the proposed rules All problems in the row indicate relationships in which A_1 is more likely given B_1 than given B_2. However, an individual using the cell a strategy would judge only the first problem as such a relationship (cell a is the largest of the cells). A person using the a versus b strategy would accurately judge the first two problems in the row, but would say that A_1 given B_1 is as likely as A_1 given B_2 in the third problem (4-4), and that A_1 was less likely given B_1 than B_2 in the last problem (2-12). The sum of diagonals rule would result in the correct judgment of the first three problems, but would say that A_1 was as likely to occur with B_1 as with B_2 on the last problem (2+10)-(12+0). A subject using the conditional probability rule should accurately judge all of the first row problems. An individual's solution pattern on the problem set would index the strategy he or she is using. Table 2b identifies the solution pattern congruent with each strategy type. The probability of matching these judgment patterns by chance alone is .11 for cell a, .04 for a versus b, .01 for sum of diagonal and .005 for the conditional probability pattern.

In two experiments, Shaklee and Tucker (1980) employed this diagnostic approach to identify judgment rules of 10th grade and college subjects. Subjects judged relationships in three problems for each proposed strategy type. Each problem consisted of 24 instances in which event states were defined for two events. Problems were set in contexts of everyday events (e.g., cake rises or falls at high or low temperature; plants healthy or not healthy which do or do not receive plant food). Subjects' performance indicated general conformity to the strategy set. Congruence with the cell a strategy pattern was frequent among the high school subjects (17%) but rare in the college sample (1%). Response patterns matched that of the a versus b strategy for 19% of the college sample (use of this strategy was not tested among the high school subjects).

Table 2

A) Cell frequencies used for each problem type

Cell a Problems			a versus b Problems			Sum of Diagonal Problems			Conditional Probability Problems		
	B_1	B_2		B_1	B_2		B_1	B_2		B_1	B_2
A_1	11	4	A_1	4	1	A_1	4	4	A_1	2	12
A_2	1	8	A_2	3	10	A_2	1	15	A_2	0	10
	B_1	B_2		B_1	B_2		B_1	B_2		B_1	B_2
A_1	6	6	A_1	4	4	A_1	9	5	A_1	1	5
A_2	6	6	A_2	8	8	A_2	7	3	A_2	3	15
	B_1	B_2		B_1	B_2		B_1	B_2		B_1	B_2
A_1	1	8	A_1	4	11	A_1	4	4	A_1	12	2
A_2	11	4	A_2	8	1	A_2	15	1	A_2	10	0

B) Strategy use and resultant patterns of problem accuracy.
(+ = accurate, 0 = inaccurate)

Subject Strategy Type	Problem Strategy Type: Cell a	a versus b	Sum of Diagonals	Conditional Probability
Conditional Probabilities	+	+	+	+
Sum of Diagonals	+	+	+	0
a versus b	+	+	0	0
Cell a	+	0	0	0
Strategy 0	0	0	0	0

Judgment patterns were congruent with the conditional probability strategy for 17% of the high school subjects and 29% of the college sample. In each experiment, the modal response pattern conformed to that of the sum of diagonals rule (34% of the college subjects, 41% of the high school subjects). Thus, overwhelmingly, subjects demonstrated at least some sophistication about appropriate covariation judgment. However; the optimal judgment rule was used by a minority of subjects in the two samples.

These initial investigations demonstrate the general success of our rule diagnostic approach. Subject judgment patterns indicated strong intraindividual consistency in rule use. Furthermore, the variety of rules evident in these results suggest that characterization of group judgment by any single rule would be inappropriate. As with all rule modeling, congruence with a rule pattern requires cautious interpretation. That is, a solution pattern conforming to one of the predicted patterns could be the product of an alternative rule which produces judgments isomorphic with the proposed rule. However, congruence with a given pattern does clearly identify the other proposed models as poor characterizations of the judgment rule. At the same time, obtained judgment patterns severely limit the pool of viable alternative models.

This rule index offers an informative method for the study of development in judgments of event contingencies. Particularly useful is the possibility of identifying specific judgment rules which might be precursors of more mature judgment competence. The steps in our strategy hierarchy may represent a developing sequence of increasingly sophisticated rule use. In fact, two of our proposed strategies, cell a and a versus b, are specifically identified by Inhelder and Piaget (1958) as characteristic of younger adolescents. The two investigators suggested that younger subjects would fail to appreciate the relevence of joint nonoccurrences of the target events (contingency table cell d) in defining relationships between event states, our cell a strategy. It was also suggested that these subjects might compare this frequency of event cooccurrences with the frequency with which one of the events occurs without the other one (contingency table cell b), our a versus b strategy. The sum of diagonals strategy was believed to develop in later adolescence, at the formal operational stage of development. Our rule diagnostic approach should allow us to track such shifts in strategy use.

Shaklee and Mims (1981) tested college subjects and children in 4th, 7th, and 10th grade on the diagnostic problem set. Again, results indicated a close congruence between actual and predicted judgment patterns. A significant developmental trend demonstrated shifts toward the use of increasingly accurate rules between the 4th to 10th grade age span. College subjects' judgments were not significantly different from those of 10th graders. Judgment patterns matched the a versus b strategy for sizable groups of subjects at all ages (21% of the college subjects, 23% of 10th graders, 25% of 7th graders, 29% of 4th graders). Sum of diagonals patterns were rare among 4th graders (17%), but common among the older subjects (38% of college subjects, 50% of 10th graders, 50% of 7th graders). Conditional probability judgment patterns were rare until the 10th grade (0% of 4th graders, 4% of 7th graders, 27% of 10th graders, and 38% of college subjects). Few people at any age level evidenced cell a judgment patterns.

Results of this developmental study suggest that development of covariation judgment may be best conceptualized as a series of approximations to optimal rule use. Early rules may afford better-than-chance performance although they are restricted in utility. With increasing age, subjects shift to more generally accurate rules. However, even among mature subjects, optimal rule use is evidenced by a minority of subjects.

A final question of concern is the stability of judgment patterns across judgment conditions. Our research indicates that close congruence between subjects' judgments and those predicted by our rules is maintained across a variety of conditions. In past work, we've varied the form in which the frequency information was presented, using individual data instances pictured on 5 x 8 cards (Experiment 1, Shaklee & Tucker, 1980), and sets of data instances pictured in a 2 x 2 table, (second experiment, Shaklee & Tucker, 1980; Shaklee & Mims, 1981). Most recently with college subjects we've presented the frequency information in numerical form. In all cases 85%-90% of subject records conformed to one of the proposed rules. Additionally, we've manipulated question form (Shaklee & Tucker, 1980, experiment 2) in testing rule use. In one experiment, some subjects were asked about the association between the events (e.g., does plant health tend to be associated with getting plant food, not getting plant food, or is there no relationship between the two?), while other subjects were asked about the likelihood of an outcome given the two possible states of the other variable (e.g., were plants who got plant food more likely, less likely, or equally likely to get well as plants who received no plant food). Subjects' judgments indicated that accuracy was higher in the latter response condition, but judgments were equally likely to match the rule patterns in the two conditions.

Finally, we conducted a pair of experiments to test rule use of college students making contingency judgments under memory load conditions (Shaklee & Mims, 1982). Since everyday covariation judgment must rely on recall of past frequency information, we were interested in rule use under more comparable conditions. Frequency information was presented in slides, each of which showed one combination of event states on the two variables. The instances in a given covariation problem were shown sequentially to subjects. In one condition, subjects tabulated frequencies as the slides are shown. In a memory condition, they estimated the frequencies after all of the instances had been shown. All subjects were asked to judge the contingency between the events once the slide sequence was shown. Subjects in the memory condition were significantly poorer at frequency estimates and also used simpler, less accurate rules than subjects in the no-memory condition. In a second experiment, subjects asked to remember distractor information in addition to the event frequency information showed more inaccurate estimates of event frequency information and use of simpler judgment strategies than subjects in a condition comparable to our prior memory condition. These two experiments indicate that covariation judgments under memory load conditions are substantially worse than those of subjects free of such memory demands.

In sum, the data from several studies indicates that a carefully structured problem set can be profitably used to indicate strategies underlying judgments of covariations between events. Such judgments are particularly interesting since they build so directly on the basic mathematical understanding of ratios and fractions. That is, people

making covariation judgments should be comparing two conditional probabilities, each of which is a ratio between two frequencies. Our evidence indicates that substantial use of such a strategy doesn't occur until the 10th grade, and then by only a minority of the subjects. This evidence is congruent with other research indicating that problems in application of ratio concepts are common among adults as well as children (Karplus & Peterson, 1970; Kurtz & Karplus, 1979; Capon & Kuhn, 1979).

Our results have further implications for statistical reasoning as a key link between math and science. Given the key role of covariation assessment in causal judgment, the evidence suggests that naive causal judgment may suffer from serious biases. That is, use of less-than-optimal judgment rules may result in erroneous perceptions of relationships between actually independent events, or failure to note relationships between events which are, in fact, related. The data further indicate that judgment problems may begin at 4th grade, when children begin to evidence reliable strategy use. Such limited rules should be particularly problematic as children enter the more advanced scientific training programs of junior high and high school. Children may make progress in rule use during those adolescent years, but biased judgment patterns persist for the majority of people even at the college years.

While the evidence clearly identifies strategy limitations among most subjects tested, those strategies may be subject to remediation. In fact, one of our previous experiments (Shaklee & Tucker, 1980) indicates that training may indeed improve performance. This experiment incorporated two types of training: concept training and sort instructions. Half of the subjects in this experiment began their sessions with a discussion of event covariations, citing covariates common to everyday life and clarifying variations in direction and strength of relationship. Comparison subjects received no such instruction. Crossed with this manipulation were instructions to sort the data instances (presented as decks of 5 x 8 cards) into a 2 x 2 matrix. Comparison subjects were not so instructed. Although sort instructions had no significant effect (half of the subjects knew to sort the data without being instructed to do so), concept training did significantly improve judgment accuracy. While the evidence indicates that training may be effective in mediating covariation judgment, the finding is somewhat general. More informative would be an approach which develops interventions specific to strategy levels.

Further evidence of the potential effectiveness of training such judgments comes from related work in probability judgment. Research by several investigators indicates that training improves probability judgments among children from 1st grade through 6th grade (Ojeman, Maxey, & Snider, 1965a & b; Shepler, 1969; Dunlap, 1980). Since covariation judgment is a comparison between probabilities, this research bolstered our expectation that training would be effective in improving rule use in covariation judgment as well.

Grant Supported Research

This program of research included a sequence of experiments designed to examine the effects of training on covariation judgment. That series began with studies to identify subjects' own understandings of their rules and sources of individual differences in rule use. The remaining experiments focus on questions about the trainability of those judgments.

Since all experiments employed the same rule analytic approach, the strategy index will first be described in detail. Discussion of specific experiments will follow.

Rule Analytic Instrument

Problems. Each subject's judgment strategy is identified through his or her solution to 12 different covariation problems, each set in the context of everyday events. Twelve different problem contents were developed, each of which consists of a set of observations picturing one of two states for two potentially related everyday events. Three problems picture bakery products which either rose or fell in association with the presence or absence of yeast, baking powder, or "special ingredient." In three other problems, plants are pictured as healthy or sick as a possible function of presence or absence of plant food, bug spray, or "special plant medicine." In three problems people (or animals) are pictured as sick or healthy as a possible function of presence or absence of a shot, liquid medicine, or a pill. The remaining three problems picture a possible association between space creatures' moods (happy/sad) and the presence or absence of one of three weather conditions (snow, fog, or rain).

For each problem, data instances are pictured in a 2 x 2 table. Example frequencies used are listed in Tables 2a and 3. Tabled frequencies indicate one noncontingent and two contingent relationships for each strategy problem type. Direction of relationship (A_1 more likely given B_1, B_2 or no difference) is counterbalanced across subjects for each problem content.

Each problem is introduced with a paragraph describing a context in which several observations were made on two potentially related variables. Subjects are asked to look at the pictured information and identify the relative likelihood of one of the events when the second event was either present or absent. An example problem:

> Spacemen from Earth landed on another planet and found creatures called the block-heads. They wanted to see what block-heads were like, so they watched them closely. Every Saturday they would look outside to check the weather and see how the block-heads were doing. Sometimes it was snowing and sometimes it was not. Sometimes the block-heads were happy and sometimes they were not. In the picture you will see how many times each of these things happened together. The picture indicates that when it was snowing block-heads were:
>
> (circle one) a) more likely to be happy than
> b) just as likely to be happy as
> c) less likely to be happy than
>
> when it was not snowing.

A similar paragraph and response form was developed for each problem content. In each case, subjects indicate whether A_1 given B_1 was more likely, just as likely, or less likely than A_1 given B_2.

Problems are grouped in blocks including a problem of each of the four strategy types. Order of problems within blocks is random. The three problem blocks are then sequenced in one of two orders and stapled together to form a problem booklet.

Instructions. Instructions introduce the concept of covariation in the context of "things which tend to go together." Real world examples are given of positive relationships (i.e., tall people are more likely to be heavy than short people), negative relationships (i.e., it is less likely to rain when the sun is shining than when it is cloudy), and unrelated events (i.e., a green truck is just as likely to run out of gas as a red truck). Subjects are told that they will do some problems about hypothetical events that may or may not tend to occur together. A sample problem is used to explain stimulus materials and problem format. Subjects progress through the problems at their own pace and are encouraged to use scratch paper if needed.

Scoring. An individual subject's performance on the four problem strategy types determines his or her rule categorization. Subjects are said to have passed criterion for any given problem strategy type if they accurately judge at least two of the three problems of that type. Table 2b lists the judgment patterns congruent with each of the proposed strategies. Subjects comparing conditional probabilities should pass criterion on all problem types. Subjects using the sum of diagonals strategy should fail on conditional proabability problems alone. Subjects comparing cell a and b frequencies should judge cell a and a versus b problems correctly, but fail sum of diagonal and conditional probability problems. The cell a strategy should result in correct judgment on cell a problems alone. The probability of meeting each of these criteria by chance alone is .11 for cell a, .04 for a versus b, .01 for sum of diagonal, and .005 for the conditional probability strategy. People who pass no criteria are labeled strategy 0. Records matching none of these patterns are labeled unclassifiable. Since problems are structured hierarchically (see Table 2b) subject's solution patterns should conform to a Guttman scale. Past evidence in our own research has indicated close congruence to the Guttman pattern (coefficient of reproducability = .93 - .97).

Experiment 1: Methods of Assessing Judgment Strategies
(See Appendix A for a detailed description of this experiment)

Thus far, we've developed our own characterization of children's rule use based on their performance on our problem set. We further evaluate those rules on the basis of the generality of their solution accuracy. A useful supplement to our own categorization and evaluation of these schemes would be information about each subject's own understanding and evaluation of his or her rule use.

The most direct approach to such information would be to interview the subjects about the strategies they're using. Interview responses could then be compared to performance patterns to note the degree of congruence between the two. Subjects' verbalizations may or may not match their performance patterns. In fact, a variety of research in psychology suggests that subjects' explanations about their judgment rules frequently do not match actual performance patterns. In development research in particular, young children's poor verbal skills may hinder

their account of systematic judgment bases. Thus, verbal accounts frequently underestimate judgment competence in research with children (Brainerd, 1973; Bullock & Gelman, 1979; Goldberg, 1966). Research with adults, on the other hand, indicates that subjects' explanations often overestimate judgment sophistication. Both expert and nonexpert judges (Goldberg, 1968; Nisbett & Wilson, 1977) describe themselves as using complex rules that bear little resemblance to the more simple patterns of their actual performance.

In order to investigate these relationships college subjects were tested with the rule analytic problem set described above. After each covariation judgment, subjects were asked to rate their confidence in the accuracy of their judgments. Once the problem set was completed, subjects were asked to explain how they solved the last problem in the set (stated strategy), to choose which of our proposed four rules was most like theirs (model choice), and to identify which of the four rules was the best one (best strategy). Each subject was tested and interviewed individually.

Results showed that problem difficulty level differed as a function of problem type, with mean accuracy decreasing as one moves up the problem hierarchy from cell a through conditional probability problem types (see Table 2, Appendix A for means). This pattern of problem difficulty replicates that seen in all of our previous studies. Subjects' confidence in their judgment accuracy also decreased as problem difficulty increased, indicating that subjects show at least some insight into the limits of their judgment rules. Males were more accurate than females in covariation judgment, although there were no sex differences in confidence ratings.

Judgment-based strategy classifications were determined as described above. Most frequently occurring were judgment patterns congruent with a versus b and conditional probability rules (36.2% and 31.9% of the samples respectively). Cell-a and sum of diagonals classifications were less common (5.2% and 15.5% respectively). Males showed use of more sophisticated strategies than females (see Table 3, Appendix A) in a pattern parallel to that found for judgment accuracy.

Subjects' interview responses were compared to judgment-based rule classifications. Correlations were significant between judgment-based and both stated strategy ($r = .58$) and model choice ($r = .45$) measures. However, examination of subject classifications shows that judgment-explanation agreement was substantially higher for conditional probability subjects (97%) than for the other strategy groups (24% for other groups combined), suggesting that some subjects knew more about what they were doing than others. Subject's choices of the best rule were found to be reliably more sophisticated than their descriptions of their own strategy (by model choice measure).

Overall, these results suggest that self-report may be a weak database for research on covariation judgment. In particular, self-report may be a poor method for diagnosing sources of error in covariation judgment. Our finding of strategy classification differences in self-report accuracy are somewhat ironic from an educational point of view. That is, the students best able to report their problem solution methods would be those who are most accurate in judgment and, hence, need help the least.

Experiment 2: Predictors of Rule Use Among College Students

Our consistent evidence throughout all of our work indicates that most subjects of a given age use a systematic rule, but that those

different rules vary substantially in sophistication and likely accuracy. We undertook a second major study with college students to try to identify sources of these individual differences in rule use. As a problem in applied mathematics, we expected that sophistication of strategy use may be related to background, aptitude, and/or interest in math. Subjects in this experiment made covariation judgments on the 12 problem sets described previously, then responded to several questions about their math background and interest (see Appendix B for complete list of questions). Questionnaire items were grouped into summary variables, including mathematics background (questions 1, 3, 8 and 17), advisors' attitudes toward the subject's participation in math courses (questions 6 and 11), the subject's self-rated mathematical ability (questions 2, 4 and 9), interest in mathematics (question 14), and perceived usefulness of mathematics (question 15). In addition,, subjects' ACT scores were obtained as potential predictors. Table 1 in Appendix B lists the intercorrelations between these variables and subject's overall judgment accuracy and strategy use. We were quite surprised to see that none of these variables were reliably related to covariation judgment. In view of the large sample size (97 females, 89 males) we regard low power to be an unlikely account of these null effects. Rather, these findings suggest that users of less accurate rules have the math background and competence for more advanced rule use. Even our most advanced rule may require math ability simple enough to be in the competence of nearly all college level students (i.e., comparison of two ratios). Those judges, however, may differ in their ability to apply that knowledge to the judgment at hand.

We also took this opportunity to further investigate our previous findings of sex differences in rule use. Again, we found that males were reliably more accurate than females in covariation judgment and that they tended to use more advanced judgment rules. Males also reported more extensive math backgrounds than females. This suggested that the sex difference in strategy use may be a function of differential math training. However, we found that the sex differences in covariation judgment were preserved even when math background was used as a covariate.

Experiment 3: Modifying Procedure For Use With Younger Subjects
(See Appendix C, Experiment 1, for detailed discussion of this study)

Our previous developmental study (Shaklee & Mims, 1981) was generally successful in characterizing systematic strategy use by subjects at the various ages tested. The one exception to this claim is the fourth grade sample, where nearly half of the children were not categorizable as using any of our proposed strategies. These unsystematic strategy users were labeled Strategy 0 (failed to pass any problem types: 21% of the sample) or unclassifiable (passed strategy types in an unpredicted pattern: 25% of the sample). Given such high rates of unsystematic judgment we must consider the possibility that these children were simply confused by the procedure. Stimuli and terminology suitable for the older children in the study may have been over the heads of these youngest subjects. If so, the fourth graders may not have been able to demonstrate the strategies within their competencies.

In view of this concern, we modified the procedure considerably to make it more appropriate for use with younger children. Our modifications centered on two major aspects. First, we wondered if these young children

were confused by the tables indicating event frequencies. As a result, we expanded our introduction of the tabled information on two practice problems, discussing the contents of each table cell and checking comprehension by asking the children to point to table cells with particular event-state pairings. We also wondered if our covariation judgment question was excessively complex syntactically. Thus, in our revised procedure, we modified the question to read (using the blockhead problem cited earlier):

The picture indicates that blockheads were more likely to be happy:

a) when it was snowing
b) when it was not snowing
c) no difference

These modifications were made to make our problems more comprehensible to younger children. It turns out that we outdid ourselves in this respect. Testing a new sample of children, nearly all of our subjects were classifiable by one of our rules in the fourth grade, and a majority of children showed systematic rule use in the second and third grades. Overwhelmingly, these subjects were classified as using the a versus b rule (see Table, 3, Appendix C). Thus, these results indicate that systematic rule use is clearly within the competence of fourth grade children, and is also common among second and third grade children. In light of these findings this modified procedure was deemed more appropriate for use with subjects in the elementary school years.

Experiment 4: Eliciting Reliable Rule Use
(See Appendix C, Experiment 2 for detailed discussion of this experiment)

Our modified procedure indicates that reliable rule use is common at an earlier age than our previous evidence indicated, but we still see that judgments are frequently unsystematic among second grade children. As a result, we were concerned about the origins of systematic rule us in judging covariation between events. Training paradigms are commonly used by psychologists to identify sources of developmental trends. If one can identify a training approach which leads an individual to show reliable rule use. contents of that training approach may indicate key aspects of knowledge that result in the natural acquisition of the rule. Of course, successful training indicates only one sufficient model of the natural developmental process. The real life transition may follow some alternative sufficient process.

We turned our attention to identifying origins of reliable rule use among first and second grade children. We chose not to train children in use of the cell a rule since it so rarely occurred naturally. Instead, we developed training programs designed to elicit use of the a versus b rule.

Our training approach stemmed from our suspicion that the judgment question itself focused children's attention on cells a and b of the contingency table. Asked if plants are more likely to be healthy when they get bug spray or when they don't get bug spray, a subject may look at those two event conjunctions (i.e., healthy plants-bug spray; healthy plants-no bug spray). We thought of this as a problem of attention direction. This was the reasoning behind our Attention only condition,

where, on a set of 6 training problems, the experimenter asked the subject to point to the event combinations specifically mentioned in the question and to count the number of cases in each of the two cells. Subjects then made their covariation judgment. Subjects had mastered this technique by the end of the training problems.

A subject may also fail to use the a versus b rule because he or she misses the comparative aspect of the question i.e., which is more likely. A second group of subjects were given the Attention instructions on the training problems and, in addition, were specifically asked which of the two cells had more cases in it. Subjects then made their covariation judgments. Subjects also mastered this technique by the end of the training problems. This group is the Attention-plus-more training group.

A final group is a no-training control group, who judged the same 6 problems but were given no special instructions.

Subjects included children in first and second grade. All subjects were pretested to establish initial rule use. Unclassifiable, Strategy 0 and cell a judges were included in the paradigm. Subjects were randomly assigned to on of the three conditions.

Subject fatigue prevented an immediate posttest of training effects. However, all subjects did return a week later for a delayed posttest. Subject's performance at that time is illustrated in Table 5 of Appendix C. Rates of improvement were at the same low level for Attention-only and control subjects. This failure of Attention-only instructions may imply that subjects at this age already know how to find the relevant cells. However, the Attention-plus-more training did result in reliable improvement at thedelayed posttest. Thus, we see that the comparative aspect of the judgment may be a key obstacle to natural use of this simple rule by young subjects.

Experiment 5: Eliciting Sum of Diagonals Rule Use

Having discovered that young children can use the a versus b rule, we next attempted to elicit use of more advanced rules from older subjects. Our first approach was to train subjects to use the sum of diagonals strategy. This strategy is built on the notion that some event combinations confirm a particular relationship between events and that some combinations disconfirm that rule. For example, if bug spray is good for plants, we should see many cases of healthy plants with bug spray and unhealthy plants without bug spray. Healthy plants without bug spray and unhealthy plants with bug spray would be exceptions to the relationship. Sum of diagonals training taught subjects that cells a and d were good examples of a positive relationship and that cells b and c were exceptions to the rule. Subjects learned that the reverse was true for negative relationships. Subjects practiced pointing to the cells with good examples and those with exceptions to the rule on each of 6 training problems. Subjects also counted the number of cases in cells a and b and in cells b and c for the training problems. These subjects then made their covariation judgments. A group of control subjects made covariation judgments on the same problems without the benefit of training. Training effects were measured in an immediate posttest and in a delayed test one week later. Subjects in the experiment were 4th, 5th, 7th and 8th grade children whose pretest performance showed use of cell a and a versus b rules.

The results of this training experiment are shown in Appendix D, Table 2. Note that unclassifiable posttest subjects were not included in the analyses. Trained subjects were significantly more likely to show use of the sum of diagonals rule both at the immediate and at the delayed posttest. This evidence indicates that subjects can indeed show improved rule use with a relatively simple training procedure. These training procedures were similarly effective among the younger and older subjects in the sample. Our training in confirming and disconfirming cases not only yielded better accuracy, but those judgments also conformed to the pattern predicted by the sum of diagonals rule. This suggests that this reasoning may well underly the natural acquisition of this rule in children's development. At a minimum, these training effects identify one sufficient model of this developmental process.

Experiment 6: Eliciting Use of the Conditional Probability Rule

Although all of our proposed rules may produce better-than-chance accuracy in covariation judgment, the conditional probability rule will correctly judge any covariation relationship. As a result, it is a matter of considerable educational significance to investigate the trainability of this rule. In view of the low incidence of use of this optimal rule at all ages tested, we should be especially motivated to find ways to improve judgment accuracy.

Our evidence thus far indicates that the conditional probability rule is the most difficult rule to train subjects to use. The subject population for this study has included seventh and eighth grade children who pretest as using the a versus b or sum of diagonals rules. Our first training approach simply taught subjects to identify the components of the relevant conditional probabilities. For example, on a problem about the effects of special plant food subjects were asked to point to the plants that got special food and count how many were there. They were then asked how many of these were healthy. In the same way, they pointed to the plants that did not get special food and noted how many were healthy. They then answered the covariation judgment question for the problem. Subjects were corrected if they made errors in identifying the components of the conditional probabilities, but received no feedback as to the accuracy of their covariation judgment. This procedure was repeated for 6 training problems. We call this condition Components training. Our evidence shows that subjects who received this training were no more likely to show use of the conditional probability rule than no-training control subjects at either immediate or delayed (one week) posttest. The training was similarly ineffective for subjects who pretested as using either the a versus b or sum of diagonals rules.

In view of these results, we amplified our training to make it much more explicit about how to combine the components of the conditional probability into two ratios, and how to make comparisons between them. Subjects then made their covariation judgments for the problem. Incorrect responses at any point were corrected including the covariation judgment itself. This procedure was repeated for 6 training problems. We call this condition the Ratio-comparison conditon. Again, subjects were junior high students who pretested as using the a versus b or sum of diagonals rules. Results of this study show that sum of diagonals subjects given Ratio-comparison training are no better than control subjects in judgment at immediate or delayed (one week) posttest.

However, a versus b subjects do show reliable improvement compared to no-training controls. It is ironic for conditional probability training to be effective for a versus b but not sum of diagonals subjects. However, examination of the data offers a ready explanation. That is, training of a versus b subjects was as likely to elicit use of the sum of diagonals rule as the conditional probability rule. Of course, subjects who pretest at sum of diagonals rule use would not be considered to have improved if they posttest as using the sum of diagonals rule. Since our "conditional probability" training seems to be as effective in eliciting use of sum of diagonals as conditional probability rule, a versus b subjects are more likely to show improvement as a result of the training.

We remain surprised that the conditional probability rule is so resistant to training. Junior high school subjects are acquainted with fractions and one might expect that comparisons of ratios would be within their competencies. However, teachers do report that fractions are a difficult subject for many students at this age. In fact, other studies find that people show problems in comparing ratios through the adult years (e.g., Capon & Kuhn, 1979; Kurtz & Karplus, 1979). However, we continue to believe that people can be trained to use the conditional probability strategy and wonder if we aren't overtaxing the mathematical competencies of junior high school children. As a result, we intend to try our Ratio-comparison training with college age subjects.

Other related research

During the period of grant support we have been involved with a couple of other projects relevant to the issue of covariation judgment. In this section, we would like to describe those studies briefly.

Thus far, our covariation judgment work has focused on judgment strategy as a determinant of judgment accuracy. However, other aspects of the judgment context may contribute to relative accuracy as well. In particular, we pursued a series of studies on the processing of event frequency information as a source of problems in covariation judgment. Our past paradigm has pictured event-state pairings in a 2 x 2 table, thus minimizing the opportunity for error in assessing the frequencies of each outcome pair. However, other information formats which occur in real life settings may produce more errors in judging events. In particular, event-state pairings occur with equal salience in our tabled format but may not be equally salient in the real world. For example, the case where two events are present (AB) may be more readily noted than a case in which the two events are absent ($\bar{A}\bar{B}$). If this were the case, a judge may reach an erroneous conclusion about the event pairings which have occurred, resulting in judgment errors by even the best of decision rules.

To test these ideas, we developed a paradigm in which a response (tapping a wire in a radio)-outcome (radio buzzes) relationship was represented in a time line (see Appendix E for a detailed description of this work). Although information about all event-state pairs was available in this format, this representation did seem to preserve the relative salience of event-state pairs that would occur with real world events (i.e., intervals with taps or buzzes more salient than those with neither tap nor buzz). Judgments in the time line representation were reliably more biased (especially for negative contingencies) than those in either

a tabled or broken time line format. These studies suggested to us that conclusions about event sequences may vary as a function of stimulus presentation conditions and may contribute to relative accuracy of covariation judgment.

Most recently, we've been looking at information sampling strategies used by subjects to test covariation and/or causal relationship. Our past research has presented subjects with information about the relative occurrences of events and asked them to draw a conclusion about the depicted relationship. However, if subjects themselves wish to test a hypothesis about an event relationship, what information would they seek?

A common pattern found in related research is a tendency to restrict one's sample to only a subset of the potentially relevant information (e.g., Shaklee & Fischhoff, 1982). In our case, a subject asked whether an outcome is associated with one of two event states might prefer to sample information about one of those event states, thereby gathering less information about the other event-state.

Such a sampling bias would have differential impact on judgment accuracy with each of our covariation judgment rules. For instance, the sum of diagonals rule is an accurate estimate of a relationship only when the two alternative states of at least one of the two events occurs equally often. The problem is easiest to demonstrate if the rule is reconceptualized in terms of its mathematical equivalent, (a-c)-(b-d). Thus, an individual compares the difference between the cells in the first column of a contingency table with the difference between cells in the second column. Those differences are only comparable estimates of likelihood if the column totals are equal; otherwise, the same sized difference represents a larger proportion of total instances for the minority event than for the majority event. The problem becomes more extreme as the difference in column totals increases, making the sum of diagonals rule an increasingly inaccurate estimate of differential likelihood. The a versus rule has the same weakness in addition to other problems. The conditional probability rule is the only strategy that will support accurate judgments with biased sampling. However, even a conditional probability rule will not handle the most extreme of sampling biases. That is, if an individual only gathers data under one event state and never samples information about the alternative state, covariation judgment cannot be at better than a chance level of accuracy. Such an individual will only know one of the two probabilities relevant to the comparison. Through preferential sampling of alternative states, people may actually generate difficult covariation problems from relationships that would otherwise be simple to evaluate.

We've developed a paradigm for investigating information search strategies used in testing hypotheses about events. Subjects were presented with a large envelope containing the universe of observations about two potentially related events on another planet. The large envelope was introduced with a description of a potential relationship between the behavior of some space creature (e.g., sleep/awake) and time of day (daytime/nighttime). It contained two smaller envelopes, each of which contained observations of the creature's behavior under one of the conditions. Thus, subjects had one envelope of daytime observations and one envelope of nighttime observations. Each observation was pictured on a 3" x 5" card in the envelope. Subjects were to select a total of

24 observations from the two envelopes in such a way as to best test a hypothesis about the events. Each hypothesis stated an association between the time of day and one behavioral state (e.g., being awake is associated with nighttime). Subjects recorded their 24 observations on a record sheet, and concluded that the hypothesis was either true or false. Subjects judged 3 such problems, including a positive, negative, and a zero relationship. These subjects also judged our strategy diagnostic problem set. Our intiial sample included third grade, seventh grade, and college subjects. We defined a subject as biased in sampling if the mean absolute differences in samples of the two event states across the three problems was greater than or equal to 8. (The range of these means could be from 0-24.) Sampling tendencies indicate that biased sampling was common at all ages tested (50% of 3rd graders, 37% of 7th graders, 32% of college subjects.) When one looks closer at the nature of the sampling bias, the overwhelming majority of cases are those in which subjects sampled solely from one envelope (day or night), ignoring information about the alternative event state. As noted earlier, this is an extent of bias that even the conditional probability strategy cannot accurately evaluate. One cannot compare two conditional probabilties with no information about one of those probabilities. Accuracy cannot be at better than chance levels under these circumstances. In overview, this information sampling paradigm does show substantial differential sampling among subjects from third grade through college age. In view of the extremity of the bias, sampling patterns such as these would be devastating to accuracy of covariation and causal judgments alike.

Overview of findings

NIE support has allowed us to investigate a variety of questions about a common form of statistical reasoning, covariation judgment. Our past work had indicated that use of systematic but simple rules began in the fourth grade and that subjects used more sophisticated rules with increasing age. Our recent research supplements this research in several important ways.

First, we found that a modified testing procedure results in spontaneous use of systematic rules at an earlier age (i.e., 2nd-4th grade) than our previous study would indicate. In addition, we found that a simple training procedure would reliably elicit use of the a versus b rule in the first and second grades. This would suggest to us that elementary school children have important competencies in understanding probabilistic relationships and may indicate that science or math demonstrations of probabilistic relationships would be suitable for children in the early primary grades.

Secondly, we find that a more advanced rule (sum of diagonals) can be acquired in the later elementary school grades with a simple training procedure. Contents of that procedure may suggest approaches which should be similarly effective in improving judgments about event covariations in classroom demonstrations in these school years. However, we find that the conditional probability rule is not easily trained in junior high children by the methods we tried. One interpretation of this finding would be that training in use of this optimal rule might be better delayed until the high school, or even college years. However, our evidence also indicates the importance of training students in these

advanced rules. That is, our past research indicates that many subjects don't spontaneously progress beyond use of the simplest, least accurate rules even in a college population.

Finally, our related research offers us evidence on non-strategic sources of covariation judgment errors. One study series shows that errors in tabulating event frequencies may be an important source of judgment bias. A second paradigm shows that subjects may look at only a subset of the information relevant to making a covariation judgment. Each of these problems would result in inaccurate judgments of event relationships by even the best of decision rules. These findings identify other aspects of information processing that may need to be incorporated into curricula designed to improve judgments of probabilistic relationships between events.

References

Adi, H., Karplus, R., Lawson, S., & Pulos, R. Intellectual development beyond elementary school VI: Correlational reasoning. School Science and Mathematics, 1978, 78, 675-683.

Brainerd, C. Judgments and explanations as criteria for the presence of cognitive structures. Psychological Bulletin, 1973, 79, 172-179.

Bullock, M. & Gelman, R. Preschool children's assumptions about cause and effect: Temporal ordering. Child Development, 1979, 50, 89-96.

Cambridge Conference on Correlation of Math and Science in the Schools. Goals for the correlation of elementary science and mathematics. Boston: Houghlin Mifflin, 1969.

Capon, N. & Kuhn, D. Logical reasoning in the supermarket: Adult female's use of a proportional reasoning strategy in an everyday context. Developmental Psychology, 1979, 15, 450-452.

Dunlap, L. First grade children's understanding of the concept of probability. Unpublished doctoral dissertation, University of Iowa, 1980.

Fischbein, E. The intuitive sources of probabilistic thinking in children. Hingham, MA: R. Reidel Publishing Company, 1975.

Goldberg, L. Simple models or simple processes? American Psychologist, 1968, 23, 483-496.

Goldberg, S. Probability judgments by preschool children: Task conditions and performance. Child Development, 1966, 37, 157-167.

Harvey, J. An immodest proposal for teaching statistics and probability. School Science and Mathematics, 1975, 75, 129-144.

Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

Inhelder, B., & Piaget, J. The growth of logical thinking from childhood to adolescence. New York: Basic Books, 1958.

Jenkins, H., & Ward, W. Judgment of contingency between responses and outcomes. Psychological Monographs, 1965, 79, 1-17.

Karplus, R., & Peterson, R. Intellectual development beyond elementary school II: Ratio, a survey. School Science and Mathematics, 1970, 70, 813-820.

Kelley, H. Attribution theory in social psychology. In D. Levine (Ed.), Nebraska Symposium on Motivation (Vol. 15). Lincoln, Nebraska: University of Nebraska Press, 1967.

Kurtz, R., & Karplus, R. Intellectual development beyond elementary school VII: Teaching for proportional reasoning. School Science and Mathematics, 1979, 79, 387-398.

Michotte, A. The perception of causality. London: Methuen, 1963.

Niemark, E. Longitudinal development of formal operations thought. Genetic Psychology Monographs, 1975, 91, 171-225.

Nisbett, R., & Wilson, T. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Ojeman, R., Maxey, E., & Snider, B. Effects of guided learning experiences in developing probability concepts at the fifth grade level. Perceptual and Motor Skills, 65, 21, 415-427, a.

Ojeman, R., Maxey, E., & Snider, B. The effect of a program of guided learning experiences in developing probability concepts at the third grade level. Journal of Experimental Education, 1965, 33, 321-330, b.

Piaget, J., & Inhelder, B. The origin of the idea of chance in children New York: Norton, 1975.

Seggie, J., & Endersby, H. The empirical implications of Piaget's concept of correlation. Australian Journal of Psychology, 1972, 24, 3-8.

Shaklee, H., Tucker, D. A rule analysis of judgments of covariation between events. Memory & Cognition, 1980, 8, 459-467.

Shaklee, H., & Mims, M. Development of rule use in judgments of covariation between events. Child Development, 1981, 52, 317-325.

Shepler, J. A study of the parts of the development of a unit in probability and statistics for elementary school. Research and Development Center for Cognitive Learning, University of Wisconsin, Madison, November, 1969.

Smedslund, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173.

Methods of Assessing Strategies for Judging Covariation Between Events

Harriet Shaklee and Laurie Hall

University of Iowa

Running head: Judging Event Covariations

in press, Journal of Educational Psychology

Abstract

Past research indicates poor agreement about strategies people use to assess covariation between events. This research investigates method of assessment as one possible source of this low consensus. A set of problems was developed in such a way that different judgment rules would produce different decisions about the relationships between events. College subjects judged these problems, then were asked to explain their judgment strategy. In addition, they were shown model strategies and asked to choose the one like their own strategy and the model that would be the best strategy. Subjects whose judgments indicated use of the most sophisticated strategy were quite accurate in reporting their judgment rules. Subjects using the less accurate rules most commonly reported using strategies which could not have produced the obtained pattern of problem solutions. These findings suggest that self-report is a weak basis for conclusions about sources of error in covariation judgment.

Statistical concepts represent one prime area for application of mathematical training. In particular, statistics are necessary for identifying predictability in an environment where relationships are frequently probabilistic (y is more likely when x is present) rather than deterministic (y always occurs when x is present). Problems such as these are common in identifying regularities in scientific phenomena, and in everyday contexts as well. In this respect, statistics provide a key link between basic mathematical concepts and central aspects of scientific and everyday problem solving. As an area for application of mathematical training, research on statistical reasoning may also be informative about children's and adult's abilities to apply their mathematical skills appropriately.

The focus of existing research in this area has been on probability judgments (e.g., Piaget & Inhelder, 1975; Fischbein, 1975; Yost, Siegel & Andrews, 1962). A statistical judgment common to reasoning about cause-effect relationships builds on probability assessments of this sort. An individual investigating the relationship between potential cause x and effect y would compare the likelihood of y occurring when x is present $P(y/x)$ with the likelihood that y occurs without x $P(y/\bar{x})$. The two events are independent if these conditional probabilities are equal; nonindependence is indicated by any difference. The comparison is made to identify contingency or covariation between events. Scientific procedure and statistical analyses testify to the key role of covariation analysis in professional practice. Although not sufficient for causal inference, covariation is a necessary condition between causes and events. Thus, covariation analysis may identify the set of possible causes of an event.

Many psychologists further assert that everyday causal judgment is similarly based on a covariation analysis (e.g., Michotte, 1963; Inhelder & Piaget, 1958; Kelley, 1967; Heider, 1958). That is, people search for likely explanations of everyday events by identifying event covariates. Thus, competence in covariation judgment may determine a person's adequacy in identifying real world cause-effect relationships.

In fact, a variety of investigators have found that adolescent and adult subjects show little competence in identifying event covariations (Niemark, 1975; Smedslund, 1963; Jenkins & Ward, 1965; Adi, Karplus. Lawson, & Pulos, 1978). While the evidence indicates that covariation judgments are often erroneous, those judgments may be rule-governed nonetheless. Several different rules have been proposed by past investigators as possible judgment strategies. These rules are discussed in terms of possible relationships between two events (A and B), each of which occurs in one of two states (1 and 2).

Least sophisticated of the proposed strategies is judgment according to the frequency with which the target events cooccur (A_1B_1, cell a in a traditionally labeled contingency table) failing to consider the other event-state pairings (A_1B_2, A_2B_1, A_2B_2) in defining the relationship. A subject using this strategy would identify a positive relationship between A_1 and B_1 if cell a frequency were the largest of the contingency table cells, a negative relationship if it were the smallest (cell a strategy). This strategy is identified by Inhelder and Piaget (1958) as common among younger adolescents. Smedslund (1963) and Nisbett and Ross (1980) suggest that the strategy is typical among adults as well. The strategy does consider some

relevant information and may result in better-than-chance performance. However, the rule considers only a limited portion of the information that defines the relationship and would result in erroneous judgment of many relationships.

A second possible approach would compare the number of times target events A_1 and B_1 cocccur with the times A_1 occurs with B_2 (comparison of frequencies in contingency table cells a and b; strategy a versus b). This strategy is also identified by Inhelder and Piaget (1958) as a precursor of matur judgment. Again this strategy considers some of the relevant information and may result in accurate judgment of many event contingencies. However, failure to consider frequencies in cells c and d (event combinations A_2B_1 and A_2B_2) would be a particularly costly error when the direction of that frequency difference is the same as the difference between cells a and b.

A much improved approach would be the strategy defined by Inhelder and Piaget (1958) as characteristic of formal operational thinking. Specifically, covariation would be defined by comparing frequencies of events confirming (cells a and d) and disconfirming (cells b and c) the relationship. Thus, the rule would compare the sums of the diagonal cells in a contingency table (sum of diagonals strategy). Jenkins and Ward (1965), however, point out that this strategy has its limits as well. Specifically, the rule is an effective index only when the two states of at least one of the variables occur equally often. Otherwise, a correlation may be indicated when, in fact, independence is the case.

Instead, Jenkins and Ward (1965) suggest that covariation is more appropriately evaluated by comparing the probability of event A_1 given

event B_1 $P(A_1/B_1)$ with the probability of A_1 given that B_2 has occurred $P(A_1/B_2)$. This is equivalent to a comparison of the frequency ratio in contingency table cells $\frac{a}{a+c}$ with that in cells $\frac{b}{b+d}$. By definition, independence is indicated by equivalence between these conditional probabilities; nonindependence is indicated by any difference (conditional probability strategy). This strategy should result in accurate judgment of any contingency problem.

Thus, four alternative strategies have been proposed to account for subjects' judgment patterns. Many of these rules were proposed on the basis of subjects' explanations of their judgments. For example, Smedslund's (1963) cell a strategy is based on the reports of over half of his sample that they judged the relation of symptom A and diagnosis F according to the number of AF pairings. Adi, Karplus, Lawson, and Pulos (1978) similarly categorized subjects according to their explanations. In this case, however, no subjects were classified as using a cell a strategy. Rather, subjects described themselves as using various combinations of two to four of the contingency table cells. Thus, two samples of subjects offer considerably different explanations of their judgment strategies. Two features of these studies make it hard to reconcile these differences. First, the two reports offer little information on the way the explanations were elicited. We might expect that different questions would result in different responses. Secondly, neither of the investigators report the level of agreement with which subject responses were categorized, so we know little about the reliability of the categorization schemes.

However, a more serious problem is relevant to any explanation-based strategy analysis. That is, such an approach is predicated on the assumption

that subjects are able and willing to accurately describe their bases of judgment. In fact, a variety of research in psychology suggests that this assumption may not be justified. In developmental research in particular, young children's poor verbal skills may hinder their account of systematic judgment bases. Thus, verbal accounts frequently underestimate judgment competence in research with children (e.g., Brainerd, 1973; Bullock, Gelman, & Baillargeon, in press; Goldberg, 1966). Research with adults, on the other hand, indicates that subjects' explanations often overestimate judgment sophistication. Both expert and nonexpert judges (e.g., Goldberg, 1968; Nisbett & Wilson, 1977) describe themselves as using complex rules that bear little resemblence to the simpler patterns of their actual performance. Ericsson and Simon (1980) note that relative accuracy of verbal reports may depend on the conditions under which the information is gathered. These findings would suggest that explanation-based analyses of judgment strategies should be treated with caution.

An alternative approach would be to analyze judgment strategies on the basis of subject's actual performance patterns (Ward & Jenkins, 1965; Jenkins & Ward, 1965; Shaklee & Tucker, 1980). That is, four different rules have been proposed to account for subjects' judgments of event covariations. Since different rules produce different judgments, covariation problems could be identified which would differentiate between those rules. In fact, careful structuring of a problem set should allow us to identify the specific strategy a subject is using.

A set of such problems is illustrated in Table 1a. Problems are structured hierarchically such that cell a problems are correctly solved by all strategies; strategy a versus b problems are correctly solved by a versus b, sum of diagonals, and conditional probability strategies. Sum

of diagonal problems will be accurately judged by sum of diagonal and conditional probability strategies. Conditional probability problems would be correctly solved by the conditional probability strategy alone. Solution accuracy is indexed by the direction of the judged relationship (i.e., A_1 more likely given B_1, B_2, or no difference). A subject's solution pattern on the set of problems indicates the strategy used. Problems on the first row of Table 1a illustrate judgments predicted by each of the proposed rules. All problems in the row indicate relationships in which A_1 is more likely given B_1 than given B_2. However, an individual using the cell a strategy would judge only the first problem as such a relationship (cell a is the largest of the cells). A person using the a versus b strategy would accurately judge the first two problems in the row, but would say that A_1 given B_1 is as likely as A_1 given B_2 in the third problem (2-2), and that A_1 was less likely given B_1 than B_2 in the last problem (2-12). The sum of diagonals rule would result in the correct judgment of the first three problems, but would say that A_1 was as likely to occur with B_1 as with B_2 on the last problem (2+10) - (12+0). A subject using the conditional probability rule should accurately judge all of the first row problems. Table 1b identifies the solution pattern congruent with each strategy type. The probability of matching these judgment patterns by chance alone is .11 for cell a, .04 for a versus b, .01 for sum of diagonal, and .005 for the conditional probability pattern.

In two experiments, Shaklee and Tucker (1980) employed this diagnostic approach to identify judgment rules of 10th grade and college students. Subjects judged relationships in three problems for each proposed strategy type. Each problem consisted of 24 instances in which event states were

defined for two events. Problems were set in contexts of everyday events (e.g., cake rises or falls with or without "special ingredient," plants healthy or not healthy which do or do not receive plant food). Subjects' performance indicated general conformity to the strategy set. Congruence with the cell a strategy pattern was frequent among the high school subjects (17%) but rare in the college sample (1%). Response patterns matched that of the a versus b strategy for 18% of the college sample (use of this strategy was not tested among the high school subjects). Judgment patterns were congruent with the conditional probability strategy for 17% of the high school subjects and 33% of the college sample. In each experiment, the modal response pattern conformed to that of the sum of diagonals rule (35% of the college subjects, 41% of the high school subjects). Subsequent studies demonstrated that children use increasingly sophisticated rules with increasing age in the 4th grade to college age span (Shaklee & Mims, 1981), and that adults tend to use simpler rules as the decision environment becomes more complex (Shaklee & Mims, 1982).

In sum, the data from several studies indicate that a carefully structured problem set can be profitably used to indicate strategies underlying judgments of covariations between events. Subjects in these experiments demonstrated at least some sophistication about appropriate covariation judgment, however, the optimal judgment rule was used by a minority of subjects. Such judgments are particularly interesting since they build so directly on the basic mathematical understanding of ratios and fractions. That is, people making covariation judgments should be comparing two conditional probabilities, each of which is a ratio between two frequencies. Our evidence indicates that substantial use of such a strategy does not occur until the 10th grade, and then by only a minority of subjects. This evidence is congruent with

other research indicating that problems in application of ratio concepts are common among adults as well as children (Karplus & Peterson, 1970; Kurtz & Karplus, 1979; Capon & Kuhn, 1979).

In addition, these findings conflict with the past interview-based strategy analyses. In particular, Smedslund's (1963) only commonly reported strategy, cell a, is rarely seen in the performance patterns of our subjects. In light of this conflict, a direct comparison of explanation and judgment-based strategy analyses would be profitable. By this approach, subjects would be asked to complete a diagnostic problem set, then explain their judgment bases. Comparison of classification by the two methods might show areas of systematic disagreement. In addition, interview responses offer new information in evaluating our judgment-based analysis. That is, subjects may describe themselves as using rules which may differ from any of our proposed rules, but which would produce a judgment pattern on the problem set congruent with that of one of our rules. Finally, we learn something about subjects' insight into their own reasoning. Such understanding of subjects' own impressions about their task solutions would be particularly important in any attempts to improve judgment competence. That is, training may be maximally effective when it is oriented toward the individual's own understanding of his or her rule use.

A second interest in this study is in subjects' evaluations of the adequacy of the rules they use. Those using less sophisticated rules may or may not be aware of rule limits. This study will measure judgments of rule adequacy by asking subjects to give confidence ratings as they make their judgments in the problem set. Subjects who are less confident of erroneous responses than of correct responses must be aware of their rule

limitations. In addition, subjects will be asked to identify the best rule among our set of proposed strategies.

Subjects for this experiment will be male and female college students, since our past research suggests that this age group should provide substantial numbers of a versus b, sum of diagonals and conditional probability judges. Sex of subject will be considered as a factor in the design in light of common findings of sex differences in math skills among adolescents and adults (e.g., Maccoby & Jacklin, 1974).

Method

Subjects

Subjects in the experiment were students in an introductory psychology class who participated in the experiment as one option in fulfillment of a course requirement. Subjects ranged in age from 18 to 32 years, with a mean age of 19.42. Sixty-two female and 54 male students participated.

Problems

Subjects judged a set of 12 covariation problems, structured so that each of four judgment rules would produce a distinctive judgment pattern on a problem set. Table 1a lists the actual problems used. The 12 problems include three problems for each of the four strategy types. One noncontingent and two contingent relationships are included for each strategy problem type.

Twelve different problem contents were developed, each of which consisted of a set of observations picturing one of two states for two potentially related everyday events. Three problems pictured bakery products which either rose or fell in association with the presence or absence of yeast, baking powder, or a "special ingredient." In three other problems, plants were pictured as healthy or sick as a possible function of the presence

or absence of plant food, bug spray, or a "special medicine." In three problems people or animals were pictured as sick or healthy as a possible function of the presence or absence of a shot, liquid medicine, or a pill. The remaining three problems pictured a possible association between space creatures appearing happy or sad in the presence or absence of one of three weather conditions (snow, fog, or rain).

For each problem, data instances are pictured in a 2 x 2 table. In each case, the manipulated factor (or environmental event) defined the table columns (e.g. plant food, no plant food in example below), and the outcomes defined the table rows (plants healthy, not healthy in the example below). Each problem is introduced with a paragraph describing a context in which several observations were made on two potentially related variables. Subjects were asked to look at the pictured information and to identify the relative likelihood of one of the events when the second event was either present or absent. An example problem follows:

> A plant grower had a bunch of sick plants. He gave some of them special plant food, but some plants didn't get special food. Some of the plants got better but some of them didn't. In the picture you will see how many times these things happened together. The picture indicates that plants which were given special food were:
>
> | +3 | +2 | +1 | 0 | -1 | -2 | -3 |
> |---|---|---|---|---|---|---|
> | much more likely | somewhat more likely | a bit more likely | just as likely | a bit less likely | somewhat less likely | much less likely |
>
> to get better than plants that weren't given special food. On your answer sheet write the scale number that best completes the sentence.

In addition, after each covariation judgment subjects were asked to rate their confidence as follows:

How certain are you in the accuracy of the above response?

1 2 3 4 5 7 7 8 9 10

just guessing absolutely certain

The 12 problems were grouped into problem blocks, including one problem from each strategy type. Problems within each block were arranged in a single random sequence. The three problem blocks were sequenced in a single random order. Numbers in parentheses to the left of the problems in Table 1a indicate the position of each problem in the problem sequence.

Once the problem set was completed each subject was interviewed and asked the following questions about his or her judgment:

1a. You've just completed several problems about the relationship between events. Can you tell how you solved them?

1b. (Experimenter turns to the last problem in the set - a conditional probability problem.) Can you use this problem to show me how you solved it? (strategy explanation)

2. (The participant is shown models of the strategy types while they are described.) Can you indicate, from the models presented, the strategy you used to solve the problems? (model choice)

3. Overall, which do you feel is the "best" strategy? (best strategy)

Each subject was tested and interviewed individually.

Instructions

Initial instructions introduced the subject to the concept of covariation in the context of "things that go together". Naturally occurring examples were given of positive relationships (i.e., tall people are more likely to be heavy than short people), negative relationships (i.e., it is less likely to rain when it is sunny than when it is cloudy), and unrelated events (i.e., a green truck is just as likely to run out of gas as a red truck). Subjects

were told that they would be given some problems about hypothetical events that may or may not tend to occur together. A sample problem involving the occurrence of snow as it did or did not relate to atmospheric temperature was used to explain the stimulus materials and the problem format. Each subject gave a solution to the sample problem and was invited to ask questions about the task. Subjects were allowed to progress through the problems at their own pace and were encouraged to use the scratch paper provided if they desired.

Results

Results can be grouped according to their relevance to two issues. First, subjects' performances can be characterized in terms of the accuracy of problem solutions. Confidence ratings on these problems indicate subjects' beliefs about their accuracy. Secondly, judgment strategies are identified according to subjects' solution patterns on the problem set and their responses to the interview questions.

Accuracy. Accuracy was assessed in terms of the direction of the judged relationship (i.e., A_1/B_1 more, less or equally likely than A_1/B_2). Data are analyzed in terms of the number of problems correct per problem type. Relevant means for this analysis are reported in Table 2. A sex by problem type analysis of variance shows a main effect of problem type ($F(3,342) = 164.36$, $p < .001$) with mean accuracy of 2.88 for cell a, 2.65 for a versus b, 1.47 for sum of diagonals, and 1.21 for conditional probability problems. A main effect of sex of subject was also significant ($F(1,114) = 6.67$, $p < .01$), with more problems correctly solved by males than by females. The sex by problem type interaction was also significant ($F(3,342) = 3.08$, $p = .03$), with the greatest sex differences in accuracy for the sum of diagonals and conditional probability problems (see Table 2).

A sex by problem type analysis of variance of confidence ratings showed that subjects had some insight into solution accuracy. This was reflected in a significant effect of problem type on confidence ratings, with confidence decreasing as problem difficulty increased ($F(3,342) = 25.60$, $p < .001$). Mean confidence ratings were 8.5 for cell a, 8.4 for a versus b, 7.8 for sum of diagonals, and 7.7 for conditional probability problems. Confidence judgments did not differ by sex either as a main effect or in interaction with problem type.

Strategy. Each subject's pattern of solution accuracy on problems of the four types was used to identify his or her judgment strategy. Performance patterns congruent with the four strategies are illustrated in Table 1b. A subject was said to have passed criterion on a given problem type if he or she was accurate on two or more of the three problems of that type. A conditional probability subject should pass criterion on all problem types, sum of diagonals judges should pass criterion on all problem types except the conditional probability problems. Judges using the a versus b rule should pass criterion on cell a and a versus b problems. Cell a subjects should pass cell a problems alone. Someone who passes no criteria would be labeled Strategy 0. Judgment patterns that do not match any of these predicted patterns are classified as "other." Classification by this method will be referred to as the judgment-based strategy.

Distribution of these judgment-based classifications is illustrated for each of the two sexes in Table 3. These results indicate that all subjects passed at least one criterion, indicating that they understood the stimuli and had at least a simple understanding of the judgment to be made. Most frequently occurring were judgment patterns congruent with a versus b and

conditional probability rules (36.2% and 31.9% of the samples respectively). Cell a and sum of diagonals classifications were less common (5.2% and 15.5% respectively). Judgments of 13 subjects failed to match any of our proposed patterns and were classified as "other". Table 3 also shows males as generally using more sophisticated strategies than those used by females. The distributions of the two sexes were compared by assigning each subject a number corresponding to the number of problem type criteria passed (cell a = 1, conditional probability = 4). A t test comparing males and females on strategy classification shows the sex difference in strategy use to be reliable ($t(101) = 2.60$, $p < .01$).

A final judgment-based strategy analysis compares the confidence ratings of subjects in each of the strategy classifications. A subject strategy by problem type analysis of variance showed no significant difference as a function of subject judgment strategy ($F(3,99) = 1.54$, ns). However, subject strategy did interact with problem type ($F(9,297) = 2.68$, $p < .01$). In this interaction, subjects classified as a versus b, sum of diagonals, and conditional probability judges showed parallel decreasing confidence as problem difficulty increased. However, cell a judges were least confident on a versus b problems. As in the previous analysis, confidence ratings also showed a main effect of problem type ($F(3,297) = 28.68$, $p < .001$).

Independent categorizations of subjects' strategies were based on their responses to the interview questions. First, subjects were asked to state their strategies (question 1a) and to demonstrate that strategy on a sample problem (question 1b). These two responses were considered together and coded according to whether they conformed to one of our four strategies. Two alternative responses were also common. Several subjects described themselves as using a variant of the conditional probability strategy which compared ratios of cell frequencies $\frac{a}{c}$ with cell frequencies

$\frac{b}{d}$. This strategy would produce the same judgments as our conditional probability strategy and will be labelled cell ratios. A second common response was for a subject to say that he or she had just guessed. Responses that did not match any of these categories were labelled "other". All responses were independently categorized by two coders. These two raters agreed on 89% of their ratings. Table 4 illustrates these classifications of subjects' explanations.

Once subjects had stated their strategies, they were shown a model of each of the four proposed strategies and asked to identify the one which most closely resembled their problem solving approach. This classification is referred to as model choice. Frequency of choices of the various models is shown in Table 5. Responses not represented in the strategy examples were coded as "other". Of these unclassifiable subjects, six said that they used more than one rule, and the remaining subjects said that they used some strategy not listed in the models.

Finally, subjects were asked to indicate the best strategy among the four examples. This response will be labelled best strategy. Table 6 lists frequencies of subjects' choices of each of the strategies. The group categorized as "other" includes several subjects who thought that two or more categories were equally good, some subjects who thought the cell ratio strategy was best, and some subjects who preferred some strategy not listed in the examples.

As in the judgment-based strategy classification, a subject's strategy classification on each of these three measures was converted to a scale score corresponding to the level of his or her classification in the strategy hierarchy. Since cell ratio judges should produce the same judgments as conditional probability rule users, these two rules were grouped

together in these analyses. Subjects who said that they guessed were given a score of 0. Comparisons between classification methods were made in terms of these scale scores. The unclassifiable subjects were not included in these analyses.

Correlations between the various strategy classifications indicate some congruence between methods. The correlation between judgment-based strategy classification and stated strategy is .58 ($p < .001$). Classification of subjects by the two methods is illustrated in Table 4. Comparisons between these classification systems indicate that differences between classifications by the two methods do not show a reliable direction ($t(94) < 1$, ns). A close inspection of Table 4 shows that performance-explanation congruence differed according to subjects' strategy classification. Subjects whose performance patterns showed use of a conditional probability rule were almost uniformly accurate in describing their strategies (97% of conditional probability subjects). Among the other groups combined (excluding "other") only 24% of the subjects described rules congruent with their performance patterns. A comparison of the two groups shows this difference to be reliable ($\chi^2 = 45.46$, $df = 1$, $p < .001$).

Comparison between judgment-based classification and subject's model choice also showed reliable congruence between the two methods ($r = .45$, $p < .001$). Table 5 shows classification of subjects by the two methods. Comparison between the classification methods shows that model choices were neither reliably more nor less sophisticated than their judgment-based strategy classification ($t(98) < 1$, ns). The correlation between the strategy explanation and model choice measure indicates some agreement between these two self-report measures ($r = .53$, $p < .001$) with the subject classifications neither better nor worse by the two methods ($t(99) < 1$, ns).

Finally, subjects' selection of best strategy was compared to their classifications by other methods. Model choice and best strategy used the same multiple choice method, and were thus deemed to make the best case for comparison (see Table 6 for classification by the two methods). Subjects' selections of best strategy were reliably more sophisticated than the strategy they identified as their own ($t(88) = 5.35$, $p < .001$), suggesting that subjects recognized a better way to solve the problems when one was provided. Their choices of best strategy were also more sophisticated than their judgment-based strategy classifications ($t(84) = 7.19$, $p < .001$).

Discussion

These results offer considerable evidence on relative congruence among self-report and performance-based methods of identifying strategies underlying covariation judgments. All comparisons suggest some agreement between methods, with correlations ranging from .45 - .58. Correlations at this level indicate that subjects have some insight into their judgment bases. However, closer inspection of Tables 4 and 5 indicate that some subjects show considerably better insight than others. In particular, conditional probability subjects (judgment-based classification) are impressively accurate, with 97% describing a conditional probsbility (or cell ratio) strategy in their strategy explanation, and 84% selecting that strategy in the model choice measure. In sharp contrast, all other subject groups show poor congruence between the performance-based and self-report measures, with 24% agreement between judgment and explanation measures, 25% agreement between judgment and model choice.

The strength of our judgment-based classification system is our ability to evaluate whether a stated rule would produce the obtained judgment pattern.

A close inspection of Table 4 illustrates this comparison. For example, no subject with a cell a judgment pattern described him or herself as using a cell a judgment rule. Our interpretation of this difference would be ambiguous if these subjects described rules which would produce a cell a judgment pattern on the problem set. However, this was not the case. Half of these subjects said they were guessing, an approach which would yield a cell a pattern only 11 percent of the time (i.e., the chance probability of producing the pattern). The remaining subjects with cell a performance patterns said they were using cell ratios, a strategy which would result in a conditional probability judgment pattern. Subjects showing a versus b patterns also showed poor insight into rule use, with 11 of 42 classifiable subjects describing themselves as using rules which should produce more errors than they actually showed, and 11 subjects describing strategies which should have produced more accurate records than actually obtained. Most of the subjects whose judgment performance indicated sum of diagonals strategy use described strategies that would produce conditional probability judgment patterns. Several subjects described themselves as comparing cells $\frac{a}{c}$ with $\frac{b}{d}$, a strategy which would mimic a conditional probability strategy on the problem set. However, it is interesting to note that only one of the subjects who said they were using cell ratios produced a judgment pattern congruent with their described rule. As noted earlier, self-report and judgment pattern were congruent for conditional probability judges. In these cases we are not simply noting relative agreement between performance and explanation. Our rule diagnostic problem set also allows us to show whether subjects' self-reported rules would have produced their actual performance patterns.

One possible interpretation of poor agreement between judgment and explanation might be that subjects shifted rule use at some point in the problem set. A subject may have judged the initial problems by one strategy, but changed strategy by the end of the problem set. This individual's judgments might yield a classification according to the initial strategy, but he or she would be accurate in describing use of a different strategy to solve the last problem. In fact, some of our subjects said that they used more than one rule in response to the model choice question. This possibility may explain a few judgment-explanation discrepancies, but our rule classification system makes it unlikely as a general account. That is, a subject had to accurately judge at least two of the three problems of each strategy type to have passed criterion on that type. The problems were blocked such that one problem of each strategy type appeared in each third of the problem sequence. A subject would have to shift strategy after the eighth problem of the set to have met the criteria for his or her initial problem solution strategy in the judgment-based classification. Shifts at other points should produce judgment records that do not conform to any of our strategy patterns. These subjects would be labeled "other" and not be included in our method comparisons. In fact, such unclassifiable subjects were infrequent in this sample (11.2%).

These results show that agreement between different self-report measures is limited as well. The correlation between subjects' strategy explanation and model choice was a modest (though significant) .53. Thus, the issue is not simply one of the validity of self-report of strategy use. Method of obtaining that self-report affects subjects' responses as well.

These comparisons suggest that self-report may be a weak data-base for research on covariation judgment. We note, however, that there may be conditions

under which self-reports would be more accurate. Our subjects described their strategies after solving a series of problems. Ericsson and Simon (1980) argue that features of memory and attention might predict that reports would be erroneous under these conditions. In particular, subjects must retrieve the relevant information from long term memory in order to explain their judgment rule. Potential sources of error include problems in storing or retrieving the information from long term memory and incomplete reporting of the available information. Ericsson and Simon (1980) argue that such problems are minimized by gathering self-reports through a think aloud technique in which subjects verbalize their reasoning as they solve the problem.

Although alternative techniques may improve self-report accuracy, our method is most relevant for comparison with past research in this area. In particular, Smedslund (1963) and Adi and colleagues (1978) each asked subjects to explain their strategies after making several judgments about event covariations. Our evidence suggests that self-report of less-than-optimal strategies will be inaccurate under these circumstances.

Considering covariation judgment as a problem in applied mathematics, our findings also have implications for educational assessment. That is, self-report may be a poor method for diagnosing the sources of individual student's errors in applying ratio concepts. Our finding of strategy classification differences in self-report accuracy are somewhat ironic from an educational point of view. That is, the students best able to report their strategies would be those who need help the least. The success of a program to improve these judgments may well depend on the starting strategy of the individual involved. Our evidence indicates that student self-report is unlikely to yield an accurate diagnosis of sources of judgment error.

Our subjects do show some insight into the strengths and weaknesses of their chosen strategies. First, confidence ratings showed that subjects were less confident of their accuracy on problems where errors were high than on problems where error rates were low. Secondly, twice as many subjects selected the conditional probability rule as the best rule as were classified as using the rule in problem solutions (32 percent vs. 65 percent). One might wonder why subjects would persist in using a rule they knew was flawed. However, shifting rules requires that subjects be able to generate a better rule to use. This evidence indicates that subjects are better at recognizing good rules than at producing those rules on their own.

A final consistent finding worth noting is the sex difference in judgment accuracy and strategy use. This sex difference is not surprising in the light of much past research showing males better than females in mathematical reasoning beginning in junior high and continuing throughout adulthood (Maccoby & Jacklin, 1974). Since the conditional probability rule builds so directly on comparisons of two ratios, we might expect sex differences in this judgment as well. Our method offers the additional advantage of identifying specific strategies employed by subjects of each sex. Compared to males, females were especially unlikely to use the conditional probability rule (19.3 percent vs. 46.3 percent), preferring the simpler and less accurate a versus b rule (41.9 percent vs. 29.6 percent). This difference could have several possible sources. One likely source is simply that the two sexes came to the experiment with different training backgrounds. Other studies have found males and females to be substantially different in participation in math courses by the time they get to college (Fennema, 1977; Keeves, 1973; Hall & Shaklee, note 1, National Assessment of

Educational Progrees, 1979). Further work would be required to assess the role of differential math training in sex differences in covariation judgments.

In overview, our results indicate that subject's self-reports of covariation judgment rules show limited congruence with actual judgment patterns. Self-report was an especially poor method for identifying sources of inaccuracy in judgment patterns. Such effects of assessment method offer a ready explanation for poor agreement about strategy use in past studies of covariation judgment. These results suggest that self-report measures are weak bases for drawing conclusions about strategy use. These problems with self-report in covariation judgment accord well with other research showing poor correspondence between subjects' judgments and their explanations about those judgments.

Reference Note

1. Hall, L. and Shaklee, H. An analysis of sex differences in judgment about covariation between events. Master's Thesis, University of Iowa, 1982.

References

Adi, H., Karplus, R., Lawson, S., & Pulos, R. Intellectual development beyond elementary school VI: Correlational reasoning. School Science and Mathematics, 1978, 78, 675-683.

Brainerd, C. Judgments and explanations as criteria for the presence of cognitive structures. Psychological Bulletin, 1973, 79, 172-179.

Bullock, M., Gelman, R., & Baillargeon, R. The development of causal reasoning. In W. Friedman (Ed.), The Developmental Psychology of Time. New York: Academic Press, in press.

Capon, N., & Kuhn, D. Logical reasoning in the supermarket: Adult female's use of a proportional reasoning strategy in an everyday context. Developmental Psychology, 1979, 15, 450-452.

Ericsson, K., & Simon, H. Verbal reports as data. Psychological Review, 1980, 87, 215-251.

Fennema, E. Influences of selected cognitive, affective, and educational variables on sex-related differences in mathematics learning and studying. In L. Fox, E. Fennema, & J. Sherman (Eds.), Women and Mathematics: Research Perspectives for Change. National Institute of Education: Washington, D. C., 1977.

Fischbein, E. The intuitive sources of probabilistic thinking in children. Boston, Mass.: R. Reidel Publishing Company, 1975.

Goldberg, L. Simple models or simple processes? American Psychologist, 1968, 23, 483-496.

Goldberg, S. Probability judgments by preschool children: Task conditions and performance. Child Development, 1966, 37, 157-167.

Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

Inhelder, B., & Piaget, J. The growth of logical thinking from childhood to adolescence. New York: Basic Books, 1958.

Jenkins, H., & Ward, W. Judgment of contingency between responses and outcomes. Psychological Monographs, 1965, 79, 1-17.

Karplus, R., & Peterson, R. Intellectual development beyond elementary school II: Ratio, a survey. School Science and Mathematics, 1970, 70, 813-820.

Keeves, J. Differences between the sexes in mathematics and science courses. International Review of Education, 1973, 19, 47-62.

Kelley, H. Attribution theory in social psychology. In D. Levine (Ed.), Nebraska Symposium on Motivation (Vol. 15). Lincoln, Nebraska: University of Nebraska Press, 1967.

Kurtz, P., & Karplus, R. Intellectual development beyond elementary school VII: Teaching for proportional reasoning. School Science and Mathematics, 1979, 79, 387-398.

Maccoby, E., & Jacklin, C. The Psychology of Sex Differences. Stanford, California: Stanford University Press, 1974.

Michotte, A. The perception of causality. London: Methuen, 1963.

National Assessment of Educational Progress. Mathematical knowledge and skills. (National Institute of Education, Report No. 09-MA-02), Denver, Colorado, 1979.

Niemark, E. Longitudinal development of formal operations thought. Genetic Psychology Monographs, 1975, 91, 171-225.

Nisbett, R., & Ross, L. Human Inference: Strategies and shortcomings of social judgment. Englewood Cliffs: Prentice-Hall, 1980.

Nisbett, R., & Wilson, T. Telling more than we can know: Verbal reports on mental processes. Psychological Review, 1977, 84, 231-259.

Piaget, J., & Inhelder, B. The origin of the idea of chance in children. New York: Norton, 1975.

Shaklee, H., & Mims, M. Development of rule use in judgment of covariation between events. Child Development, 1981, 52, 317-325.

Shaklee, H., & Mims, M. Sources of error in judging event covariations: Effects of memory demands. Journal of Experimental Psychology: Learning, Memory & Cognition, 1982, 8, 208-224.

Shaklee, H., & Tucker, D. A rule analysis of judgments of covariation between events. Memory and Cognition, 1980, 8, 459-467.

Smedslund, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173.

Ward, W., & Jenkins, H. The display of information and the judgment of contingency. Canadian Journal of Psychology, 1965, 19, 231-241.

Yost, P., Sigel, A., & Andrews, T. Nonverbal probability judgments by young children. Child Development, 1962, 33, 769-780.

Footnotes

Partial support for this research was provided by NIE grant NIE-G-80-0091. Many thanks to Renee Smith for her help in collecting this data. Reprint requests should be sent to Harriet Shaklee, Department of Psychology, University of Iowa, Iowa City, Iowa 52242.

[1]We had some difficulty defining a noncontingent relationship for the sum-of-diagonals problems. The problem we included (middle problem, column 3, Table 1A) deviates slightly from independence ($P(A_1|B_1) - P(A_1|B_2) = -.06$) by the conditional-probability rule. As a result we scored responses as correct if subjects concluded that $A_1|B_1$ was either less likely or just as likely as $A_1|B_2$. The problem does discriminate appropriately between the other judgment rules. Cell-a and a-versus-b judges should say that $A_1|B_1$ is more likely than $A_1|B_2$, sum-of-diagonal [illegible] should say the two outcomes are equally likely.

Table 1

A) Cell frequencies used for each problem type

Cell a Problems

(11)

	B_1	B_2
A_1	11	2
A_2	4	7

(3)

	B_1	B_2
A_1	6	6
A_2	6	6

(5)

	B_1	B_2
A_1	2	11
A_2	7	4

a versus b Problems

(6)

	B_1	B_2
A_1	7	3
A_2	2	12

(9)

	B_1	B_2
A_1	3	3
A_2	9	9

(4)

	B_1	B_2
A_1	4	11
A_2	8	1

Sum of Diagonal Problems

(2)

	B_1	B_2
A_1	2	2
A_2	2	18

(7)

	B_1	B_2
A_1	9	7
A_2	5	3

(10)

	B_1	B_2
A_1	8	8
A_2	8	0

Conditional Probability Problems

(8)

	B_1	B_2
A_1	2	12
A_2	0	10

(12)

	B_1	B_2
A_1	1	5
A_2	3	15

(1)

	B_1	B_2
A_1	12	2
A_2	10	0

B) Strategy use and resultant patterns of problem accuracy.
(+ = accurate, 0 = inaccurate)

Subject Strategy Type	Problem Strategy Type: Cell *a*	*a* versus *b*	Sum of Diagonals	Conditional Probability
Conditional Probabilities	+	+	+	+
Sum of Diagonals	+	+	+	0
a versus *b*	+	+	0	0
Cell *a*	+	0	0	0
Strategy 0	0	0	0	0

Table 2

Mean Judgment Accuracy Per Problem Type

	cell *a*	*a* versus *b*	sum of diagonals	conditional probability	all types
females	2.81	2.64	1.23	1.00	1.90
males	2.96	2.65	1.72	1.43	2.20
all	2.88	2.65	1.47	1.21	2.05

Table 3

Judgment-Based Strategy Classifications

(percentages)

	cell a	a versus b	sum of diagonals	conditional probability	other	N
males	3.7	29.6	11.1	46.3	9.3	54
females	6.4	41.9	19.3	19.3	12.9	62
all	5.2	36.2	15.5	31.9	11.2	116

Table 4

Frequencies of Strategy Classifications by Judgment-Based And Strategy Explanation Methods

	Strategy Explanation							
Judgment Based	guess	cell *a*	*a* vs *b*	sum of diagonals	conditional probability	cell ratios	other	all
cell *a*	3	0	0	0	0	3	0	6
a versus *b*	9	2	13	2	1	8	7	42
sum of diagonals	3	0	1	1	6	6	1	18
conditional probability	0	0	0	1	35	1	0	37
other	2	0	1	0	0	10	0	13
all	17	2	15	4	42	28	8	116

Table 5

Frequencies of Strategy Classifications by Judgment-Based And Model Choice Methods

	Model Choice						
Judgment Based	guess	cell a	a versus b	sum of diagonals	conditional probability	other	all
cell a	0	1	3	1	1	0	6
a versus b	6	4	14	7	10	1	42
sum of diagonals	2	2	2	1	9	2	18
conditional probability	2	0	0	3	31	1	37
other	1	0	1	1	6	4	13
all	11	7	20	13	57	8	116

Table 6

Frequencies of Strategy Classifications by Model Choice

And Best Strategy Methods

Best Strategy	Model Choice: guess	cell *a*	*a* versus *b*	sum of diagonals	conditional probability	other	all
cell *a*	0	1	0	1	0	0	2
a versus *b*	2	0	4	1	0	0	7
sum of diagonals	1	2	4	1	1	0	9
conditional probability	6	2	10	4	49	5	76
other	2	2	2	6	7	3	22
all	11	7	20	13	57	8	116

APPENDIX B

Questionnaire and table: Predictors of covariation judgment strategy use

JUDGING RELATIONSHIPS BETWEEN EVENTS SURVEY

Please answer each of the following questions to the best of your ability. Write your responses on the accompanying answer sheet.

1. How much math did you take during grades 9 and 10?

2. Please estimate the quality of your performance in these courses?

1	2	3	4	5
poor		average		excellent

(Write the number corresponding to your estimate on your answer sheet for each scale of this type).

3. How much math did you take during grades 11-12?

4. Please estimate the quality of your performance in these courses:

1	2	3	4	5
poor		average		excellent

5. Did you ever seek guidance from a high school counselor or counselors regarding election of math courses?

6. Please indicate the general attitude of any counselors consulted regarding your election of math courses:

1	2	3	4	5
very unfavorable		neutral		very favorable

7. Please indicate the amount of influence the counselor's advice had on your election of math courses:

1	2	3	4	5
none		some influence		strong influence

8. How many college mathematics and math-related semesters have you completed thus far?

9. Please estimate the quality of your performance in these courses:

1	2	3	4	5
poor		average		excellent

10. Have you sought guidance from your college advisor regarding your election of math courses?

11. Please indicate the attitude of your advisor toward your electing math courses:

1	2	3	4	5
very unfavorable		neutral		very favorable

12. Please indicate the amount of influence your advisor's recommendations had on your election of math courses:

1	2	3	4	5
none		some influence		strong influence

13. How many math and math-related courses do you expect to take in the future?

14. Please indicate the amount of interest mathematics holds for you:

1	2	3	4	5
boring		neutral		interesting

15. Please estimate the usefulness of mathematical knowledge to your future career:

1	2	3	4	5
not at all useful		maybe useful		extremely useful

16. How many semesters of logic have you taken?

17. How many semesters of statistics or probability have you taken?

18. What is your major course of study?

19. How favorable is your mother's attitude toward your pursuing a college education?

1	2	3	4	5
very unfavorable		neutral		very favorable

20. How favorable is your father's attitude toward your pursuing a college education?

1	2	3	4	5
very unfavorable		neutral		very favorable

Thank you very much for your cooperative participation.

Table 1

Correlation coefficients and
number of observations (in parentheses)
for Questionnaire data

	MATH	ACC	STRAT	ACT-Q	ACT-C	INTRST	USE	ABIL
ACC	.16 (186)							
STRAT	.04 (161)	.91*** (161)						
ACT-Q	.37*** (186)	.18 (186)	.15 (161)					
ACT-C	.28*** (186)	.17 (186)	.16 (161)	.81*** (186)				
INTRST	.41*** (186)	.15 (186)	.12 (161)	.35*** (186)	.24** (186)			
USE	.43*** (186)	.22 (186)	.20 (161)	.23** (186)	.22* (186)	.53*** (186)		
ABIL	.32*** (104)	.09 (104)	.03 (90)	.39*** (104)	.33** (104)	.58*** (104)	.34** (186)	
ATT	.28 (54)	.12 (54)	.04 (51)	.27 (54)	.15 (54)	.32* (54)	.21 (54)	.24 (37)

* $p < .01$
** $p < .001$
*** $p < .0001$

MATH: Math background
ACC: Accuracy on covariation judgment problems
STRAT: Covariation judgment strategy
ACT-Q: Quantitative score on ACT exam
ACT-C: Combined score on ACT exam
INTRST: Interest in mathematics
USE: Usefulness of mathematics
ABIL: Self-rated math ability
ATT: Counselor's attitude toward math

Eliciting Systematic Rule Use in Coordination Judgment

Harriet Shaklee and Donald Paszek

University of Iowa

Running head: Covariation Judgment

Partial support for this research was provided by NIE grant NIE-G-80-0091. Many thanks to Ernest Jones, Renee Smith, Rick Taffe and Janet Lyness for their help with data collection. Reprint requests should be sent to Harriet Shaklee, Department of Psychology, University of Iowa, Iowa City, Iowa 52242.

Abstract

Related research suggests that children may show some simple understanding of event covariations by the early elementary school years. The present experiments use a rule analysis methodology to investigate covariation judgments of children in this age range. In Experiment 1, children in second, third and fourth grade judged covariations on 12 different covariation problems. Children's performance patterns on the problem set showed an increase in the use of systematic judgment strategies in this age range. Systematic rule users most commonly compared contingency table cells a and b in judging the event covariations. In Experiment 2, a training paradigm was employed to investigate possible origins of systematic rule use. First and second grade unsystematic, strategy 0 and cell-a children were either directed to attend to cells a and b (Attention only), were additionally offered explicit instructions to note which of the two cells had more events (Attention-plus-more) or were given no training (control). Posttest performance showed that the Attention-plus-more condition was the only treatment to reliably elicit a-versus-b rule use. It is concluded that simple covariation judgment rules can be used by children in the early elementary school years.

Covariation Judgment: Systematic Rule Use in the Early Years

Interest in children's causal reasoning has burgeoned in recent years (e.g., Siegler, 1976; Bullock, Gelman & Baillargeon, 1982). A number of theorists have suggested that identification of cause-effect relationships is grounded in covariation judgment (e.g., Inhelder & Piaget, 1958; Kelley, 1972). That is, people search for causes of events by finding event covariates. In fact, a few investigations indicate that children understand this link from an early age. For example, DiVitto and McArthur (1978) found that children as young as first grade use summarized covariation information in explaining people's behavior. Siegler and Liebert (1975), however, found that children were not influenced by event covariation until 8 or 9 years of age in their study of children's explanations of physical events. Evidence of the earliest use of event covariation in causal reasoning is provided by Shultz and Mendelson (1975), who found that 3 and 4 year old children showed a preference for covariates when choosing causes of events. Although the age trends differ in these studies, they concur in suggesting that preference for consistent covariates is an early developing pattern in children's explanations of events.

Given this evidence, understanding development in covariation judgment would be critical to understanding children's causal reasoning. However, investigations of children's abilities to make covariation judgments are rare indeed. Those few studies which do exist show a degree of consensus on how children might judge event relationships (Inhelder & Piaget, 1958; Adi, Karplus, Lawson & Pulos, 1978; Shaklee & Mims, 1981). In the basic paradigm investigators offered subjects information on the frequency of cooccurrence of alternative event states of two potentially related variables (for example, plants healthy or not healthy; plant food present or absent). Subjects were asked to identify the direction and/or strength of the relationship between the events. In each

experiment, subjects' covariation judgments and/or explanations of the judgments led the investigators to identify systematic but inaccurate rules which were precursors to the use of more mathematically sophisticated rules.

Inhelder and Piaget (1958) proposed two simple rules of covariation judgment. In the first, an individual would judge a relationship according to the frequency with which target event states cooccur (e.g., healthy plants which are given plant food in the example above, cell a of a traditionally labeled contingency table. See Table 1). A subject using this strategy would

---

Insert Table 1 here

---

identify a positive relationship between events if the cell a frequency were the largest of the contingency table cells, and a negative relationship if it were the smallest (cell-a strategy). Inhelder and Piaget (1958) identified this strategy as common among younger adolescents. Smedslund (1963) and Nisbett and Ross (1980) thought the strategy might typify adult reasoning as well.

Also proposed by Inhelder and Piaget (1958) was a second simple approach comparing the number of times the target outcome occurs with the supposed cause (or covariate) with the number of times it occurs without that cause (for example, healthy plants with plant food vs. healthy plants without plant food). This would compare contingency table cells a and b (strategy a-versus-b). This strategy was identified by Inhelder and Piaget (1958) as typical of young adolescents and was found by other investigators to be common among high school subjects as well (Adi, Karplus, Lawson and Pulos, 1978).

Inhelder and Piaget (1958) proposed a third strategy as characteristic of formal operational thinking. That is, subjects would compare frequencies of events confirming (cells a and d) and disconfirming (cells b and c) a relationship of a particular direction. This rule would compare the sums of diagonal cells in the contingency table (sum of diagonals strategy).

Finally, Jenkins and Ward (1965) propose that covariation is most accurately assessed by comparing the conditional probabilities of an event occurring given each of the alternative states of the other variable (e.g., plant health/plant food vs. plant health/no plant food). This would compare the frequency ratio in contingency table cells $\frac{a}{a + c}$ with that in cells $\frac{b}{b + d}$ (conditional probability strategy).

This analysis of possible rules may allow diagnosis of strategies actually employed by children of various ages. That is, different rules should produce different judgments on carefully constructed covariation problems. A set of such problems is illustrated in Tables 2a and 2b. Solution accuracy is indexed

---

Insert Tables 2a and 2b here

---

by the direction of the judged relationship (i.e. A_1 more likely given B_1, B_2, or no difference). Problems are structured hierarchically such that cell-a problems are correctly solved by all strategies, a-versus-b problems are accurately solved by all strategies except cell-a. Sum-of-diagonals problems are accurately judged by sum-of-diagonals and conditional probability strategies and conditional probability problems are accurately judged by the conditional probability rule (see Table 3). The probability of matching these

Insert Table 3 here

judgment patterns by chance alone is .11 for cell-a, .04 for a-versus-b, .01 for sum of diagonals, and .005 for the conditional probability pattern.

Shaklee and Mims (1981) used this rule diagnostic approach to study covariation judgment strategies used by subjects from 4th grade through college age. Subjects' judgment patterns in that age span showed a strong developmental trend, with the a-versus-b strategy evidenced by substantial numbers of subjects beginning in the fourth grade (29%), and sum of diagonals the modal strategy at 7th and 10th grade (50% of subjects). Conditional probability patterns were produced by many subjects at the 10th grade (27%) but were still used by a minority of subjects even in the college years (38%). Thus, this evidence supports previous investigators' suggestions that children may use simpler, less accurate rules as precursors to mature reasoning. However, these results deviated from previous conclusions in two notable ways. First, the commonly proposed cell-a judgment pattern was rare among subjects at any of the ages tested (0-8%). In addition, the level of mature reasoning most often fell short of the optimal judgment strategy.

These results further contrast with findings in the causal reasoning research where use of covariation information was seen in causal judgment anywhere from preschool to 8-9 years of age. Shaklee and Mims (1981), on the other hand, find that nearly half of fourth graders showed no systematic bases of covariation judgment. A look at the causal reasoning research indicates that these studies offered children a relatively easy task of covariation judgment. DiVitto and McArthur (1978), for example, summarized the covariation information for the subjects, allowing children to use the information in causal judgment when they might not be able to derive that information for themselves.

In the remaining studies (Shultz & Mendelson, 1975; Siegler & Liebert, 1975), the target event and its possible causes were either perfectly contingent or completely independent. Studies of covariation judgment, on the other hand, commonly ask for judgments about less-than-perfect relationships. This analysis would indicate that young children may evidence a very simple understanding of covariation which does not hold up well when judging relationships of intermediate strength.

A final related paradigm must also be considered in understanding children's covariation judgment. That is, one commonly employed test of probability judgment is one in which a child is shown two piles of marbles composed of different proportions of marbles of two colors. The subject is asked to indicate the pile from which he or she would rather make a blind choice in order to obtain the marble of a particular color. The judgment is formally comparable to a covariation judgment, where a subject decides if a given outcome is more likely under condition A_1 or A_2. Siegler's (1981) rule analysis of children's performance in this paradigm shows systematic rule use by a narrow majority of 5 year olds with most of those children using a rule comparable to the a-versus-b rule in covariation judgment research. By 8-9 years of age a substantial majority of children were using systematic judgment rules, with a comparison of conditional probabilities the modal response pattern in Experiment 1, a-versus-b the dominantly used rule in Experiment 2. Each experiment found a comparison of conditional probabilities to be the most common rule among 12 year olds and adults.

Thus, in contrast to covariation judgment research, Siegler found that systematic rule use in a related judgment occurs at an earlier age, culminating in use of the optimal rule by early adolescence. Siegler's (1981) findings may suggest that Shaklee

and Mims (1981) provide a conservative estimate of children's acquisition of systematic bases of covariation judgment. Causal reasoning research also indicates that some simple understanding of event covariation may be seen by the early elementary school years. Possible resolution of these differences may begin with a careful look at the covariation judgment paradigm. The reliable strategy use evidenced by older subjects clearly indicates that they understood the experimental stimuli and procedures. However, among the fourth grade sample, 25% of the subjects produced unclassifiable response patterns, and an additional 21% passed no strategy criteria at all. This high rate of unsystematic responses may indicate that a substantial group of these children were confused by the paradigm and thus, unable to demonstrate systematic rules which may be in their repertoires. If this were the case, a simplified approach should be developed to test these younger subjects.

We address the question of early covariation judgment in two ways. Experiment 1 employs a simplified paradigm to examine the development of covariation judgment rule use among young elementary school children. Once these normative trends are established our second study investigates sources of this shift to systematic rule use. In Experiment 2, we test information components which may be sufficient to elicit reliable rule use among young children.

Experiment 1

Simplification of our previous experimental procedure was accomplished in two major ways. First, we were concerned that younger subjects might not understand the stimuli represented in the 2 x 2 table. As a result, a new introduction expanded the discussion of the contents of the table, asking the subject to point to examples of each of the four possible combinations of event states in the table.

Secondly, we suspected that our previous question format might be overly complex for the younger children. The previous question asked (in the plant food example discussed above).

When they got special food, plants were

a) more likely to be healthy than

b) just as likely to be healthy as

c) less likely to be healthy than

when they didn't get special food.

A reformulated question offered simpler syntax:

Plants were more likely to get better if

a) they got the special food

b) they did not get the special food

c) no difference

We expected that this simplified question would be more appropriate to the language competencies of younger subjects. Experiment 1 also included two different problem sets in anticipation of our needs in the subsequent study.

Method

Subjects

Subjects in the experiment were respondents to an advertisement in a small town newspaper offering payment to second, third and fourth grade children for participating in a psychology experiment. The resultant sample included 37 second graders, 18 third graders, and 17 fourth graders.

Problems

Subjects judged one of two sets of 12 covariation problems, each structured to produce a distinctive pattern of solution accuracy by each of the four proposed judgment rules. In one set of problems, cell frequencies totaled 36 for each problem (set 24), in the other set, cell frequencies totaled 36 for each problem (set 36). Except for these frequency differences, the two problem sets were identical in other respects. Tables 2a and 2b show the actual problem frequencies used for the problems in each of the two sets. The 12 problems in each set included three problems for each of the four strategy types. One noncontingent (middle row Tables 2a and 2b) and two contingent relationships (top and bottom rows Tables 2a and 2b, $P(A_1/B_1) - P(A_1/B_2) = (.40$ to $.50)$

were included for each problem strategy type. Table 3 shows the pattern of solution accuracy congruent with each of the proposed rules.

Each problem was set in a concrete context of two everyday events which may or may not be related. Each individual event pairing was illustrated with a small picture showing the state of the two variables (e.g., plant sick or healthy/plant food present or absent). Three problems pictured bakery products which either rose or fell in association with the presence or absence of yeast, baking powder, or a "special ingredient". In three other problems, plants were pictured as healthy or sick as a possible function of the presence or absence of plant food, bug spray, or a "special medicine". In three problems people or animals were pictured as sick or healthy as a possible function of the presence or absence of a shot, liquid medicine, or a pill. The three remaining problems pictured a possible association between space creatures appearing happy or sad in the presence or absence of one of three weather conditions (snow, fog, or sunshine).

For each problem, data instances were organized in a 2 x 2 table. In each case, the manipulated factor (or environmental event) defined the table columns (e.g., plant food, no plant food in example below), and the outcomes defined the table rows (e.g., plants healthy, not healthy in the example below). Each problem was introduced with a paragraph describing a context in which several observations were made on two potentially related variables. Subjects were asked to look at the pictured information and to identify the relative likelihood of one of the events when the second event was either present or absent. An example problem follows:

A plant grower had a bunch of sick plants. He gave some of them special plant food, but some plants didn't get special food. Some of the plants got better but some of them didn't. In the picture you will see how many times these things happened together. The picture shows that the plants were more likely to get better if:

A. they got the special food.

B. they did not get the special food.

C. no difference (they were just as likely to get better with food as without the food).

The 12 problems were grouped into problem blocks, including one problem from each strategy type. Problems within each block were arranged in a single random sequence. The three problem blocks were sequenced in a single random order. Numbers in parentheses to the left of the problems in Tables 2a and 2b indicate the position of each problem in the problem sequence.

Procedure

Each subject was tested individually. Introductory instructions introduced the subject to the concept of covariation in the context of "things that go together". Naturally occurring examples were given of positive relationships (i.e., tall people are more likely to be heavy than short people), negative relationships (i.e., it is less likely to rain when it is sunny than when it is cloudy), and unrelated events (i.e., a green truck is just as likely to run out of gas as a red truck). Subjects were told that they would be given some problems about hypothetical events that may or may not tend to go together. Two sample problems were used to clarify the information in the 2 x 2 table. The first sample problem was read to the subject. The subject was told that

pictures in the cells showed the occurrence or nonoccurrence of the two events in the story. The experimenter then pointed out that each cell represented a different combination of the two possible events and stated what these were. The subject was asked to point to cells corresponding to specific combinations of events given by the experimenter. The experimenter explained that each picture in the cells represented one occurrence of a particular combination of events, so that the number of pictures in each cell represented the number of times that combination occurred. The experimenter then read the covariation question to the subject and asked him or her to answer it based on the events pictured in the table. It was emphasized that subjects should answer the questions based on what had occurred in each story problem and should avoid basing answers on knowledge of common everyday occurrences (for example, that it is more likely to snow when it is cold, regardless of cell frequencies). Each subject gave a solution to the problem and repeated the procedure on the second sample problem. Subjects were encouraged to ask any questions they might have about the task.

The subject then proceeded to the 12 problem set. Each of the problems in the set were read to the subject by the experimenter. Subjects were allowed to answer the problems at their own pace.

Results

Our main interest in this study was to establish trends in strategy use among these younger subjects. As a result, the analyses in this study use subject strategy classification as the dependent variable of interest. Subjects were classified for strategy use according to the method illustrated in Table 3. A subject was said to have "passed" a given problem type if he or she was accurate on two or more of the three problems of a given problem type.

A subject who met this criterion on all problem types would be classified as a conditional probability rule user, subjects who passed criteria on all types except the conditional probability problems were labeled sum-of-diagonals judges. A-versus-b judges should pass the cell-a and a-versus-b problems, but not the other problem types, cell-a rule users should pass criterion on cell-a problems alone. Subjects who passed no problem types were labeled Strategy 0; all other judgment patterns were categorized as unclassifiable. Table 4 shows the rule classifications of subjects in each of the three grades.

Insert Table 4 here

The modal classification at each of the grades was a-versus-b, with very few subjects showing evidence of use of more sophisticated rules and a few subjects at each grade with cell-a rule judgment patterns. Many subjects in the second and third grades made judgments that were not classifiable by any of our rules. Effects of grade level and problem set were examined by assigning subjects a score according to the number of problem type criteria passed. Thus, Strategy 0 subjects were assigned a score of 0, conditional probability subjects a score of 4. Unclassifiable subjects could not be clearly ranked in this way and were excluded from these analyses. Data from the remaining subjects were analyzed in an analysis of variance with subject's grade (2, 3, or 4) and problem set (24 or 36) as factors. These analyses showed a significant effect of grade, F (2,51) = 3.30, $p < .05$, with third and fourth graders similar to each other, and classified as using more advanced rules than the 2nd graders (Duncan's multiple range test, $p < .05$). Problem set effects were not significant.

Discussion

Related research in causal reasoning and probability judgment indicated that children might show some simple understanding of event covariation by early elementary school. This experiment found that a majority of children do show systematic rule use in covariation judgment by the second grade. Significant age trends also show an increase in systematic rule use with age in the second to fourth grade age span. Rule categorizations in this age range show a substantial decline in unclassifiable and Strategy 0 subjects with increasing age and an increase in a-versus-b rule use. However, use of more advanced rules was rare at all ages tested.

Comparison with Shaklee and Mims (1981) indicates that subjects did indeed show earlier competencies with our revised procedure. Nearly all fourth graders were classifiable by one of our proposed rules in the present experiment and a majority of children showed systematic rule use in the second and third grades. Overwhelmingly, these children were classified as using the a-versus-b rule. The low frequency of more sophisticated strategies is comparable to that seen in our prior research. Also, similar to our past results is the low rate of usage of the cell-a strategy. This is especially interesting, given that it is the most common of the proposed judgment strategies and was even said to be the modal strategy among adults (Smedslund, 1963; Nisbett & Ross, 1981). Our evidence finds this strategy to be rare among children as young as second grade.

These results would indicate that our prior procedures may have been unnecessarily confusing to younger subjects. Our prior and present procedures were not systematically compared in this paradigm, nor did we compare aspects of the changed procedure (e.g., instruction vs. question format) in a factorial design. As a result, we can offer little information about what aspects of the prior procedure may have been a problem. However, it is clear that we have developed a procedure suitable for use with young children. These findings

indicate that children as young as second grade use simple but systematic rules in judging event relationships.

Age trends in this paradigm show origins of rule use in covariation judgment at age levels comparable to that of researchers in causal and probabilistic reasoning. However, in one respect, these results differ from Siegler's (1981) data on children's probability judgments. In those experiments, substantial numbers of children used the conditional probability rule by 8-9 years. In fact, a comparison of conditional probabilities was the modal response pattern in this age group in one of his experiments. In contrast, none of the subjects in this experiment was classified as using the conditional probability rule and only a few used the sum of diagonals rule. Our past research (Shaklee & Tucker, 1979; Shaklee & Mims, 1982) found the conditional probability rule to be used by only a minority of subjects even at adulthood. Thus, comparability between these paradigms in terms of early rule use is not matched by performance similarity in the later years. Expressing a judgment in terms of marbles in piles elicits more advanced rule use than a question asking for a comparable decision in terms of covariations between potentially related events. One difference may be that our problems are set in contexts of events that are readily interpreted as causally related. Adi and colleagues (1978) found th subjects used simpler, less accurate rules in evaluating cause-effect relationships than in making covariation judgments on analogous problems. Evidence such as this may indicate that covariation judgment in a causal context lags behind the same judgment about non-causal relationships.

Our evidence of systematic rule use at an early age is intriguing, but equivalently interesting are the unsystematic judgments of so many age peers. That is, at second and third grades a majority of children are classified by one of our rules (59% and 61% respectively), but a substantial minority in

each grade produce unsystematic judgment patterns (19% and 39% respectively) or pass no problem type criteria (Strategy 0 = 22% of second graders). Inspection of individual subjects' judgment patterns failed to identify any alternative strategic bases of these responses. Thus some children are unsystematic in rule use at the same age as other children begin to show use of simple judgment strategies. What did these rule users know that allowed them to judge the problems in a systematic fashion? Several factors may differentiate these rule users from their unsystematic age peers.

One possibility may be that unsystematic subjects are not using the tabled frequencies at all, but rather are judging the event covariations on the basis of their prior expectations about the event relationships. For example, such children may decide that plants are more likely to be healthy when they get plant food based on their real world experience, regardless of the event frequencies in the problems they are asked to judge. Our instructions already caution subjects against making expectancy-based judgments but those instructions may be readily forgotten as the subject solves the problems.

Expectancy-based judgments may be a source of unclassifiable response patterns, but what leads others of these young subjects to adopt an a-versus-b rule? We suspected that the judgment question itself may direct children's attention to cells a and b of the contingency table. Asked if plants are more likely to be healthy when they get plant food or when they do not get plant food, a subject may look at these two event conjunctions (i.e., healthy plants-plant food, healthy plants-no plant food). A subjects must also attend to the comparative aspect of the question in order to employ the a-versus-b rule. Mastery of either the attention direction or comparative aspects of the judgment (or both) may be key competencies underlying the shift to a-versus-b rule use at these early ages.

These are plausible sources of development in covariation judgment, but their roles in the origins of systematic rule use have yet to be demonstrated. An approach often employed to model a naturally occurring developmental trend is a training paradigm. That is, one might identify a training program which teaches non-rule users the knowledge said to differentiate those subjects from rule-based age peers. Contents of a successful training procedure identify at least one sufficient model to account for the natural transition to systematic rule use.

Experiment 2

We propose to use this training strategy in Experiment 2 to investigate the origins of systematic rule use in judging event covariation. Results of Experiment 1 indicated that reliable rule use was already becoming common in the second grade sample. Thus, Experiment 2 was an attempt to train first and second grade children to use the a-versus-b rule. We chose not to train children in use of the cell-a rule since it so rarely occurred naturally.

If young children's judgments are unsystematic because they are expectation-based, this problem would best be treated by drawing children's attention to the frequency information in the tables. Thus, one training procedure directed children's attention to the frequencies involved in the a-versus-b rule, i.e., cells a and b. This was the reasoning behind the Attention-only condition, where, on a set of 6 training problems, the experimenter asked the subject to point to the event combinations specifically mentioned in the question and to count the number of cases in each of the two cells. Subjects then made their covariation judgment.

As suggested previously, a subject may also fail to use the a-versus-b rule because he or she misses the comparative aspect of the question i.e., which is more likely. A second group of subjects were given the Attention instructions on the training problems and, in addition, were specifically asked which of the two

cells had more cases in it. Subjects then made their covariation judgments. This group is the Attention-plus-more training group.

A final group is a no-training control group, who judged the same 5 problems but were given no special instructions.

All subjects were pretested to establish initial rule use. Unclassifiable, Strategy 0 and cell-a judges were included in the paradigm. Subjects were randomly assigned to one of the three conditions. Training effects were measured in a posttest given about a week after the training session. In view of their comparability in Experiment 1, problem set 24 and set 36 were problems in this experiment.

Method

Subjects

Subjects were respondents to ads in a small town newspaper offering first and second graders payment for participation in a psychology experiment. Forty-nine subjects participated in the pretest session of the experiment. However, 13 subjects were dropped from the experiment because their pretest strategy indicated that they were already using the a-versus-b (9 subjects) or a more advanced strategy (3 sum-of-diagonals subjects, 1 conditional-probability subject). The remaining 36 subjects (18 males and 18 females) included 13 unclassifiable, 17 Strategy 0, and 6 cell-a subjects. Mean age of these subjects was 7 years-6 months (range 6 years-10 months to 8 years-0 months).

Pretest

Problems and instructions on the pretest were identical to those described in Experiment 1. Half of the subjects were given problem set 24 for the pretest and set 36 for posttest, the remaining subjects were given the problem sets in the reverse sequence.

Once the problem set was completed, the experimenter determined the subject's judgment strategy in the manner described in Experiment 1.

Training

Six new problems were developed for training. These problems used cell frequencies and contents which were different from those used in the two test sets. Subjects classified as cell-a, Strategy 0, or unclassifiable were randomly assigned to one of three training conditions (12 subjects per condition).

Attention-Only. This training was designed to direct subject's attention to the two event pairings specifically mentioned in the question (i.e., cells a and b). Verbatim instructions for this condition were as follows (portions were re-phrased if necessary):

> In doing these problems, you may have had a certain way of deciding which answer you thought was right. For example, you may have thought that certain boxes and the pictures in them were important and other boxes were not important in answering the question. Or you may have compared certain boxes with each other. If one thing happened more than another thing, it may have been more likely to happen. Now we are going to see if there might be another way to solve these problems that may be better than the way you used. We will try to decide which boxes and the pictures in them are important in deciding which answer is right. I want you to think hard now about a good way to answer these problems. I'll ask you some questions to help figure out a way to decide what answer is right.
>
> (The first problem and question were read to the child.)
>
> If we wanted to decide which answer is right, it is important to look at each answer and find good examples or pictures that may show that thing happening. For example, let us suppose we wanted to see if answer A might be the right answer. Answer A says (e.g., the bugs are more likely to crawl on the leaves when it is sunny out). Could you show me which box or pictures are good examples of that? Which pictures show where the (bugs crawl on the leaves when it is sunny out)?

(Subjects should point to cell a, and were corrected if they did not. When subjects did point to cell a:)

Right. Can you tell me why? So these pictures show the (bugs crawling on the leaves when it is sunny out). This is an important box to look at in deciding if answer A is right. And how many times did that happen? So there are _____ good examples of answer A.

(The experimenter also pointed to other cells, asked or pointed out why they were not good examples.)

Now let us look at answer B, because that could also be the right answer.

(The same procedure was repeated. Subjects should point to cell b. The experimenter selected answer A and answer B to be discussed first with approximately equal frequencies. The discussion was then summarized.)

Okay, so that means that if we wanted to see if (question with answer A is read) this box (cell a) and the pictures in it would be important to look at. And we see that it happened _____ times. If we wanted to see if (question with answer B is read) this box (cell b) and the pictures in it would be important to look at. And we see that this happened _____ times. It is also possible that answer C is correct, that it didn't make any difference (if it was sunny or not, the bugs were just as likely to crawl on the leaves)

The covariation judgment question was then read to the subject and he or she made a response.

Attention-plus-more. This training condition was designed to emphasize the comparative aspect of the question, i.e., which outcome was more likely? The training builds on the Attention-only training described earlier. Subjects in this condition heard all of the instructions in the Attention-only training, and were then asked to make a direct comparison of cell a and cell b frequencies ("Which of

these two things happened more?"). The experimenter then read the covariation judgment question to the subject and he or she made a response.

Control. Subjects in this condition judged the same problems as subjects in the other groups, but were offered no training instructions.

In each training condition, the procedure described was repeated on the six training problems. Feedback (positive or negative) was not provided following the subject's answers to the covariation judgment question.

Posttest

Subject fatigue prevented an immediate posttest of training effects. However, all subjects did return approximately one week later for a delayed posttest. This posttest was administered by a second experimenter who was blind to the training condition of the subject. The experimenter first reviewed the stimulus materials and problem format by presenting one of the sample problems used in session 1. Following this, the second problem set was administered in the same manner as in session 1. Subjects were tested on the problem set (24 or 36) not judged in the pretest session. Following completion of the problem set, subjects were told the purpose of the experiment and its potential relevance to everyday causal reasoning.

Results

The first indication of the relative success of the training methods was children's performance on the 6 training problems. Subjects responded in the manner predicted by the a-versus-b rule on 43.1% of the problems in the control group, 72.2% of the problems in the Attention-only group, and 97.2% of the problems in the Attention-plus-more group. An overall analysis of variance indicates these differences to be reliable, $F(2,33) = 18.81$, $p < .001$. Pairwise comparisons indicate that each training group is significantly different from each of the other groups (Duncan's multiple range test, $p < .05$).

Effects of the training procedure are most clearly assessed by comparison of the posttest performance of subjects in the training and control conditions. These effects will be analyzed both in terms of the accuracy of subjects on the various problem types and in terms of their posttest strategy classifications.

For each subject, posttest judgment accuracy was assessed in terms of the percentage of correct judgments for each of the 4 problem types. These data were analyzed in an analysis of variance including problem type (4 levels) and subject's training condition (3 levels) as factors. This analysis indicated a significant main effect of problem type, F (3,99) = 17.22, $p < .001$, and a significant interaction between problem type and training condition, F (6,99) = 5.78, $p < .001$. As the means indicate in Table 5, Attention-plus-more subjects were

Insert Table 5 here

substantially more accurate on cell-a and a-versus-b problems than on sum of diagonals and conditional probability problems. Attention-only and control subjects' performance were similarly poor across problem types. The main effect of training condition was not significant.

Pretest and posttest strategy classifications were compared for each subject to note training effects. Judgment was said to have improved if a subject was classified as using the a-versus-b, sum of diagonals, or conditional probability strategy at posttest. Table 6 indicates the frequencies of improvement

Insert Table 6 here

of subjects in each of the three training conditions. In all cases subjects who improved were categorized as using the a-versus-b strategy.

An overall χ^2 shows these training effects to be significantly different between conditions ($\chi^2 = 11.02$, df = 2, $p < .01$). As indicated in the table, rates of improvement were at similarly low levels (25%) in the control and Attention-only conditions compared with substantial rates of improvement (83%) among Attention-plus-more subjects.

Discussion

These results offer clear evidence of the differential effectiveness of our various training conditions. First, spontaneous improvement from test to retest was rare among subjects in the control condition. This would suggest that these young subject's problems were not simply lack of familiarity with the problems.

Improvement rates were equally low in the Attention-only condition. This null effect indicates that simply directing attention to cells a and b is not sufficient to elicit a-versus-b rule use among these children. The failure of Attention-only instructions may imply that subjects at this age already know how to find the cells mentioned in the question. If this were the case control and Attention-only subjects would be essentially equivalent in knowledge state at posttest. One would also expect that the Attention-only training would be sufficient to overcome any tendency to make expectation-based judgments. That is, children's attention was repeatedly directed to the information in the table cells. Indeed, the children's improved performance on the training problems suggests that the training was successful in eliciting frequency-based judgments. However, these effects were not maintained at the posttest one week later. Of course, any null effect has at least one alternative interpretation. That is, the Attention-only training condition may have simply been ineffective in teaching the knowledge that should have been sufficient to elicit [illegible]

However, the Attention-plus-more training did result in reliable improvement at the posttest. This finding indicates that the comparative aspect of the judgment may be a key obstacle to natural use of this simple rule by young subjects. Although they may know that two cells of the table are relevant, apparently subjects this young cannot spontaneously derive a way to combine that information to make a single judgment. Our training in the "more" rule apparently offers them that information. Since this training builds on the information offered in the Attention-only condition, this effect may hinge on the combined influence of the attention direction and comparative aspects of the question. Unfortunately a "More-only" condition is logically impossible. One cannot talk about comparing cells without designating which cells are to be compared. The fact that these training effects held over a one week delay period indicates the reliability of knowledge the children acquired.

Finally, it is worth noting the specificity of our training effects. That is, all children who improved in strategy use showed use of the a-versus-b strategy. This aspect of the results indicates that subjects were not simply learning to be systematic in judgment bases. Rather, they acquired one specific judgment rule. On this problem set, use of the a-versus-b rule did not lead to an overall improvement in judgment accuracy. This is by design of the problem set. That is, a-versus-b judges should be correct on cell-a and a-versus-b problems but incorrect on the sum of diagonals and conditional problems. Thus, the successful Attention-plus-more training actually results in worse performance of half of the problems compared to the other two conditions.

These training effects offer one sufficient model of the natural process of acquiring the a-versus-b rule. That is, subjects whose attention was directed to cells a and b and who were instructed to compare the two cells

showed a-versus-b rule use. Thus, these two knowledge components may be the source of children's natural shifts to a-versus-b rule use. Of course, a sufficient process is not always a necessary one. That is, children may spontaneously discover the rule through yet another sufficient process.

These training effects may also be appreciated in a broader context. That is, research in causal reasoning indicates that some simple understanding of event covariation may begin in early elementary school (Shultz & Mendelson 1975; Siegler & Liebert, 1974). Siegler's (1981) work in probability judgment shows similar age trends in children's use of simple rules in comparing probabilities. This evidence indicates that these competencies may be shown at an even earlier age with a brief training procedure. It may be interesting to see if these improvements in covariation judgment would influence children's causal reasoning as well. This may be a domain in which to test children's ability to apply statistical concepts appropriately to related judgments.

Whether children could learn to use a more complex rule with appropriate training is a question for future research. However, the level of math involved in our other rules may preclude their use in early elementary school. The sum of diagonals rule requires a comparison of two sums, the conditional probability rule compares two ratios. These advanced arithmetic competencies are likely to be outside of the capacity of such young children.

In overview, these two studies offer new information about covariation judgment in the early elementary school years. That is, many children spontaneously show use of the a-versus-b rule as early as second grade. Children as young as first grade can be taught to use this simple rule if offered the relevant information. This training evidence offers one sufficient model of the natural acquisition of a simple rule for judging relationships between events.

References

Adi, H., Karplus, R., Lawson, S., & Pulos, R. Intellectual development beyond elementary school VI: Correlational reasoning. School Science and Mathematics, 1978, 78, 675-683.

Bullock, M., Gelman, R., & Baillargeon, R. The development of causal reasoning. In W. Friedman (Ed.), The Developmental Psychology of Time. New York: Academic Press, 1982.

DiVitto, B., & McArthur, L. Developmental differences in the use of distinctiveness, consensus, and consistency information in making causal attributions. Developmental Psychology, 1978, 14, 474-482.

Inhelder, B., & Piaget, J. The Growth of Logical Thinking from Childhood to Adolescence. New York: Basic Books, 1958.

Jenkins, H., & Ward, W. Judgment of contingency between responses and outcomes. Psychological Monographs. 1965, 79, 1-17.

Kelley, H. Attribution theory in social interaction. In E. Jones, et al., (Eds.), Attribution: Perceiving the causes of behavior. Morristown, NJ: General Learning Press, 1972.

Nisbett, R., & Ross, L. Human Inference: Strategies and Shortcomings of Social Judgment. Englewood Cliffs, NJ: Prentice-Hall, 1980.

Shaklee, H., & Mims, M. Development of rule use in judgments of covariation between events. Child Development, 1981, 52, 317-325.

Shultz, T., & Mendelson, R. The use of covariation as a principle of causal analysis. Child Development, 1975, 46, 394-399.

Siegler, R. Defining the locus of developmental difference in children's [illegible]. Journal of Experimental Psychology, 1976, 30, 512-525.

[illegible], S. Three aspects of cognitive development. Cognitive Psychology, [illegible], 4, 481-520.

Siegler, R. Developmental sequences within and between concepts. Monograph of the Society for Research in Child Development, 1981, 46, Whole No. 189.

Siegler, R., & Liebert, R. Effects of contiguity, regularity and age on children's causal inferences Developmental Psychology, 1974, 10, 574-579.

Smedslund, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173.

Footnote

[1]We had some difficulty defining a noncontingent relationship for the sum of diagonals problems. The problem we included (middle problem, column 3, Tables 2a and 2b) deviates slightly from independence ($P(A_1/B_1) - P(A_1/B_2) = -.06$, set 24, $-.03$ set 36) by the conditional probability rule. As a result we scored responses as correct if subjects concluded that A_1/B_1 was either less likely or just as likely as A_1/B_2. The problem does discriminate appropriately between the other judgment rules. Cell-a and a-versus-b judges should say that A_1/B_1 is more likely than A_1/B_2, sum of diagonal judges should say the two outcomes are equally likely.

Table 1

Contingency Table Cell Labels

	B_1	B_2
A_1	a	b
A_2	c	d

Table 2

A) Cell frequencies used for problems in problem set 24.

Cell a Problems

(11)

	B_1	B_2
A_1	11	2
A_2	4	7

(3)

	B_1	B_2
A_1	6	6
A_2	6	6

(5)

	B_1	B_2
A_1	2	11
A_2	7	4

a versus b Problems

(6)

	B_1	B_2
A_1	7	3
A_2	2	12

(9)

	B_1	B_2
A_1	3	3
A_2	9	9

(4)

	B_1	B_2
A_1	4	11
A_2	8	1

Sum of Diagonal Problems

(2)

	B_1	B_2
A_1	2	2
A_2	2	18

(7)

	B_1	B_2
A_1	9	7
A_2	5	3

(10)

	B_1	B_2
A_1	8	8
A_2	8	0

Conditional Probability Problems

(8)

	B_1	B_2
A_1	2	12
A_2	0	10

(12)

	B_1	B_2
A_1	1	5
A_2	3	15

(1)

	B_1	B_2
A_1	12	2
A_2	10	0

B) Cell frequencies used for problems in problem set 36.

Cell a Problems

(11)

	B_1	B_2
A_1	16	4
A_2	6	10

(3)

	B_1	B_2
A_1	9	9
A_2	9	9

(5)

	B_1	B_2
A_1	3	16
A_2	10	7

a versus b Problems

(6)

	B_1	B_2
A_1	11	2
A_2	7	16

(9)

	B_1	B_2
A_1	5	5
A_2	13	13

(4)

	B_1	B_2
A_1	6	18
A_2	9	3

Sum of Diagonal Problems

(2)

	B_1	B_2
A_1	4	4
A_2	4	24

(7)

	B_1	B_2
A_1	12	9
A_2	9	6

(10)

	B_1	B_2
A_1	11	11
A_2	11	3

Conditional Probability Problems

(8)

	B_1	B_2
A_1	3	18
A_2	0	15

(12)

	B_1	B_2
A_1	2	7
A_2	6	21

(1)

	B_1	B_2
A_1	18	3
A_2	15	0

Table 3

Strategy Classification Criteria

Strategy use and resultant patterns of problem accuracy.

(+ = accurate, 0 = inaccurate)

		Problem Strategy Type			
		Cell *a*	*a* versus *b*	Sum of Diagonals	Conditional Probability
Subject Strategy Type	Conditional Probability	+	+	+	+
	Sum of Diagonals	+	+	+	0
	a versus *b*	+	+	0	0
	Cell *a*	+	0	0	0
	Strategy 0	0	0	0	0

Table 4

Experiment 1

Rule classifications of subjects by Grade (percentages)

Grade	Strategy Classification						
	Unclassifiable	Strategy 0	Cell-a	a-versus-b	Sum of Diagonals	Conditional Probability	N
2	19	22	16	40	3	0	37
3	39	0	11	44	6	0	18
4	6	0	18	71	6	0	17

Table 5

Experiment 2

Mean percent correct for each problem type

Training condition	Problem Type: Cell-a	a-versus-b	Sum of Diagonals	Conditional Probability	All
Attention-plus-more	83.3	80.6	8.3	5.5	44.4
Attention-only	55.4	44.3	27.8	33.3	40.2
Control	52.8	38.8	44.4	24.8	40.2
All	63.8	54.6	26.8	21.2	41.6

Table 6

Effects of a-versus-b training on posttest performance

	Improved	Didn't Improve	Total
Control	3	9	12
Attention Only	3	9	12
Attention plus more	10	2	12
Total	16	20	36

Training for Improved Covariation Judgment

Harriet Shaklee, Laurie Hall, and Don Paszek

University of Iowa

Paper presented to the Psychonomics Society, November, 1982, Minneapolis, Minnesota.

This research was supported by a National Institute of Education Grant G-80-0091 awarded to the first author. Many thanks to Paul Holt and Nancy Oatken for assistance in data collection.

A variety of theorists have suggested that covariation judgment may be a key element in causal reasoning. That is, people may find likely causes of an event by searching for covariates of that event. If causal and covariation judgment are interlinked in this way, then accuracy of covariation judgment may set an upper limit to an individual's competence at causal reasoning.

Evidence from our own investigations indicates that people show wide individual differences in competence at covariation judgment. In particular, a majority of adults employ rules which may lead to better than chance accuracy, but which result in systematic errors on same event relationships. We've focused our investigation on four strategies which might account for subjects' judgment patterns. Each of these strategies will be discussed in terms of the four cells of a 2 x 2 contingency table, labeled cells a, b, c, and d in a left to right, top to bottom sequence. One commonly proposed strategy is to judge a relationship according to the number of times the target event states co-occur, cell-a of the contingency table. We term this strategy the cell-a strategy. A second approach might compare the number of times the target event occurs with its supposed cause with the number of times that event occurs without that possible cause. This strategy would compare frequencies in contingency table cells a and b, a strategy we call a-versus-b. A third strategy might compare the number of events confirming a relationship of target event and supposed cause with the number of events which would disconfirm such a relationship. This strategy would compare the sum of frequencies in cells a and d with that of cells b + c, a strategy we term sum of diagonals ((a + d) - b + c)). Finally, a mathematically sophisticated approach would compare the probability of target event

given the supposed cause with the probability of the event when that cause was absent. We call this strategy the conditional probability strategy and is the only one of our strategies which will always produce correct judgments of event covariations.

Thus, we propose four different judgment rules varying in complexity and likely accuracy. Since different rules should produce different judgments, we can construct a problem set where each solution strategy produces a unique solution pattern. A sample of such problems is illustrated in Table 1a. Problems are structured hierarchically such that cell-a problems are accurately judged by all rules; a-versus-b problems should be correctly judged by all but cell-a judges. Sum-of-diagonals problems should be accurately judged by sum-of-diagonals and conditional-probability. problems should be accurately judged by the conditional probability rule, alone. Accuracy of judgment is indexed by the direction of the judged relationship. For example, a-versus-b judges should judge the conditional probability problem in Table 1a as a case in which A_1 is less likely given B_1, than given B_2 (2-12). Sum of diagonals judges should judge the two events as unrelated (2 + 10 = 0 + 12) and conditional probability judges should see A_1 as more likely given B_1 than given B_2 (2/2 vs. 12/22). A subjects' strategy is indexed by the accuracy pattern on a 12 problem set, including 3 problems of each of the problem strategy type. Table 1b indicates judgment accuracy predicted by each of the proposed rules. Subjects who pass no problem types are labeled Strategy 0. All other patterns not represented in the table would be labeled unclassifiable. We've looked at rule use in this way in several experiments involving subjects from 4th grade through college age. Problems in these experiments are set in the context of concrete events which could

be related. Frequency information is represented in pictorial format in a 2 x 2 table. Subjects are asked about the relative likelihood of an outcome given the two alternative states of the other variables.

Our past evidence indicates a strong developmental trend in the 4th grade to college age span. The modal strategy at 4th grade was the a-versus-b rule, although Strategy 0 and unclassifiable judges were also common. The sum of diagonals rule was used by a substantial group of subjects in our 7th and 10th grade samples. The conditional probability rule was used by a substantial minority of subjects in tenth grade and college. The cell-a rule was rare at all ages tested. Thus, subjects used increasingly sophisticated rules with increasing age. However, the optimal conditional probability rule was used by a minority of subjects even at college age.

Having discovered these developmental trends, our current efforts are trying to account for those trends. That is, what knowledge differences between these age groups may be implicated in the differences in rule use. A common approach to the problem is to develop a training method which is effective in eliciting use of more advanced rules. Contents of those effective interventions allow us to identify one sufficient account of naturally occurring developmental trends. Effective training programs may also be of pragmatic value in improving covariation judgment.

Our first concern was with the many fourth graders who didn't match any of our proposed rules. Given the number of such subjects, we have to consider the possibility that these children were confused by some aspect of our method and were unable to demonstrate their true competencies. Our approach was to elaborate our instructions to insure that the children understood the tabled stimuli and to reformulate the covariation question

in a syntax more appropriate for younger children.

These modifications were made to make our problems more comprehensible to younger children. It turns out that we outdid ourselves in this respect. Testing a new sample of children, nearly all of our subjects were classifiable by one of our rules in the fourth grade, and a majority of children showed systematic rule use in the second and third grades. Overwhelmingly, these subjects were classified as using the a-versus-b rule. Unclassifiable and Strategy 0 judgment patterns were predominant among first and second grade children. As a result, this population was the target age for an attempt to elicit use of a simple judgment rule. Thus, the first experiment I'll describe is an attempt to train 7 year old subjects to use the a-versus-b rule. We opted not to train children in use of the cell-a rule since it so rarely occurred naturally.

Our training approach stemmed from our suspicion that the judgment question itself focused children's attention on cells a and b of the contingency table. Asked if plants are more likely to be healthy when they get bug spray or when they don't get bug spray, a subject may look at those two event conjunctions (i.e. healthy plants-bug spray; healthy plants-no bug spray). We thought of this as a problem of attention direction. This was the reasoning behind our attention only condition, where, on a set of 6 training problems, the experimenter asked the subject to point to the event combinations specifically mentioned in the question and to count the number of cases in each of the two cells. Subjects then made their covariation judgment. Subjects had mastered this technique by the end of the training problems.

A subject may also fail to use the a-versus-b rule because he or she misses the comparison aspect of the question i.e., which is more likely. A second group of subjects were given the Attention instructions

on the training problems and, in addition, were specifically asked which of the two cells had more cases in it. Subjects then made their covariation judgments. Subjects also mastered this technique by the end of the training problems. This group is the Attention-plus-More training group.

A final group is a no training control group, who judged the same 6 problems but were given no special instructions.

All subjects were pretested to establish initial rule use. Unclassifiable, Strategy 0 and cell-a judges were included in the paradigm. Subjects were randomly assigned to one of the three conditions.

Subject fatigue prevented an immediate posttest of training effects. However, all subjects did return a week later for a delayed posttest. Subject's performance at that time is illustrated in Table 2 of your handout. As you can see, rates of improvement were at the same low level for Attention-only and control subjects. This failure of Attention-only instructions may imply that subjects at this age already know how to find the relevant cells. However, the Attention-plus-More training did result in reliable improvement at the delayed posttest. Thus, we see that the comparative aspect of the judgment may be a key obstacle to natural use of this simple rule by young subjects.

Having discovered that young children could use this simple rule, we next attempted to elicit use of more advanced rules from older subjects. Our first approach was to train subjects to use the sum-of-diagonals strategy. This strategy is built on the notion that some event combinations confirm a particular relationship between events and that some combinations disconfirm that rule. For example, if bug spray is good for plants, we should see many cases of healthy plants with bug spray and unhealthy plants without bug spray. Healthy plants without bug spray and unhealthy

plants with bug spray would be exceptions to the relationship. Sum-of-diagonals training taught subjects that cells a + d were good examples of a positive relationship and that cells b + c were exceptions to the rule. Subjects learned that the reverse was true for negative relationships. Subjects practiced pointing to the cells with good examples and those with exceptions to the rule on each of 6 training problems. Subjects also counted the number of cases in cells a + b and in cells b + c for the training problems. These subjects then made their covariation judgments. A group of control subjects made covariation judgments on the same problems without the benefit of training. Training effects were measured in an immediate posttest and in a delayed test one week later. Subjects in the experiment were 4th, 5th, 7th and 8th grade children whose pretest performance showed use of cell-a and a-versus-b rules.

The results of this training experiment are shown in Table 3. Note that unclassifiable posttest subjects were not included in the analyses. Trained subjects were significantly more likely to show use of the sum-of-diagonals rule both at the immediate and at the delayed posttest. This evidence indicates that subjects can indeed show improved rule use with a relatively simple training procedure. These training procedures were similarly effective among the younger and older subjects in the sample. Our training in confirming and disconfirming cases not only yielded better accuracy, but those judgments also conformed to the pattern predicted by the sum-of-diagonals rule. This suggests that this reasoning may well underly the natural acquisition of this rule in children's development. At a minimum, these training effects identify one sufficient model of this developmental process.

Our final efforts at training are looking at what it takes to elicit use of the optimal conditional probability rule among junior high aged subjects. Thus far it looks like our training efforts are successful. This set of training studies suggests that subjects at all ages may show problems in covariation judgment but that those problems are not irremediable. Our evidence suggests that relatively simple training efforts can elicit use of more sophisticated and more accurate judgment rules.

Training for Improved Covariation Judgment

Harriet Shaklee, Laurie Hall, and Don Paszek

University of Iowa

Paper presented to the Psychonomics Society, November, 1982, Minneapolis, Minnesota.

Table 1

A) Sample covariation problems

Cell a Problem

	B_1	B_2
A_1	11	4
A_2	1	8

a versus b Problem

	B_1	B_2
A_1	4	4
A_2	3	16

Sum of Diagonal Problem

	B_1	B_2
A_1	4	4
A_2	21	15

Conditional Probability Problem

	B_1	B_2
A_1	2	12
A_2	0	10

B) Strategy use and resultant patterns of problem accuracy. (+ = accurate, 0 = inaccurate)

Subject Strategy Type	Problem Strategy Type: Cell a	a versus b	Sum of Diagonals	Conditional Probability
Conditional Probabilities	+	+	+	+
Sum of Diagonals	+	+	+	0
a versus b	+	+	0	0
Cell a	+	0	0	0
Strategy 0	0	0	0	0

This research was supported by a National Institute of Education Grant G-80-0091 awarded to the first author. Many thanks to Paul Holt and Nancy Oetken for assistance in data collection.

Table 2

Effects of a-versus-b Training on Delayed Posttest performance of 7 year old children

	Improved	Didn't Improve	Total
Control	3	9	12
Attention Only	3	9	12
Attention plus more	10	2	12
Total	10	20	36

$\chi^2 = 11.02$, df = 2, $p < .01$.

Table 3

Effects of Sum-of-Diagonals Training

on Immediate and Delayed Posttest

performance on 4th-8th grade children

	Immediate Posttest			Delayed Posttest			
	Improved	Didn't Improve	Unclassifiable	Improved	Didn't Improve	Unclassifiable	N
Control	4	17	2	5	14	4	23
Training	15	6	8	21	6	2	29
Total	19	23	10	26	20	6	52

$\chi^2 = 9.6$, df = 1, $p < .01$

$\chi^2 = 9.87$, df = 1, $p < .01$

Judging Response-Outcome Relations: The Role of Response-Outcome Contingency, Outcome Probability, and Method of Information Presentation

Edward A. Wasserman and Harriet Shaklee

The University of Iowa

Running head: Judging Response-Outcome Relations

Abstract

A series of four experiments investigated college students' judgments of interevent contingency. Subjects were asked to judge the effect of a discrete response (tapping a wire) on the occurrence of a brief outcome (a radio's buzzing). Pairings of the possible event-state combinations (response-outcome, response-no outcome, no response-outcome, no response-no outcome) were presented in a summary table (Experiments 2 and 4), in an unbroken time line (Experiments 1, 2, and 4), or in a broken time line format (Experiment 3). Subjects judged the extent to which the response caused the outcome or prevented it from occurring. Across all methods of information presentation, judgments were a positive function of response-outcome contingency and outcome probability. In the unbroken time line condition, judgments of negative response-outcome contingencies were less extreme than judgments of equivalent positive contingencies. This asymmetry was smaller in the broken time line condition and in those conditions where subjects were encouraged to segment an unbroken time line into discrete response-outcome units. Finally, judgments of positive and negative relationships were generally symmetrical in the summary table condition. Relative to the two time line portrayals, summary table judgments were also less influenced by the overall probability of outcome occurrence. These judgment differences among format conditions suggest that, depending on the method of information presentation, subjects differently partition event sequences into discrete event pairings. The segmenting of continuous event streams may be an important factor in the accuracy of everyday judgments of interevent contingency.

And now remains,
That we find out the cause of this effect,
Or rather say the cause of this defect,
For this effect defective comes by cause.

W. Shakespeare; Hamlet, II, ii

Students of behavior both before and after Shakespeare have been interested in causal perception. Most noteworthy was D. Hume (1739) who proposed a set of conditions which were conducive to cause-effect impressions. Hume's insights into the psychology of causation have helped to shape the direction of subsequent research and theory in the area.

Also important have been discussions of causal perception from comparative and developmental perspectives. C. L. Morgan (1893, 1894) concluded on the basis of extremely limited evidence that human adults, but not children and animals, can perceive the relationship between events. More systematic data led Inhelder and Piaget (1958) to propose a stagewise unfolding of the human's conception of interevent correlation or contingency as the individual develops from child to adult.

Subsequent investigations into the perception of interevent relations have not yielded evidence that is consistently favorable to the developmental and evolutionary speculations of Morgan and of Inhelder and Piaget. Nor is the evidence particularly supportive of modern theories, which posit a virtual identity between humans' and animals' perceptions and the actual interevent contingencies that prevail in their environments (e.g., Heider, 1958; Kelley, 1967; Mackintosh, 1974; Rescorla, 1978).

In the basic human judgment paradigm, subjects are given information about the frequency of pairings of alternative states (e.g., presence and absence) of two events (e.g., plant food and plant health); they can then be asked to judge the direction and magnitude of the relationship between the events. In many of these experiments, adults do not accurately judge the correlation between two binary variables (see Crocker, 1981 for a review).

Despite these negative results, other work has been more successful in showing that adults can accurately judge interevent relations under some circumstances (e.g., Allan & Jenkins, 1980; Alloy & Abramson, 1979; Seggie, 1975; Seggie & Endersby, 1972; Shaklee & Tucker, 1980). Nevertheless, many factors have been suggested over the past 20 years which may contribute to distortions in the perception of correlation.

Investigators have found that the accuracy of correlational judgments depends on the sign of the relationship being judged. In particular, Erlick and Mills (1967) found that subjects judged negative correlations as closer to zero than positive correlations of equal magnitudes. Also common is the result that subjects find contingencies of zero to be especially difficult to identify. For example, Seggie (1975) reported that subjects were accurate in their judgments of contingent relationships, but were error-prone in judging noncontingent relationships (also see Allan, 1980; Allan & Jenkins, 1980). Alloy and Abramson (1979) replicated this pattern of differential accuracy in nondepressed subjects, but found that depressed adults judged noncontingent problems closer to zero than did nondepressed subjects.

One must, however, be cautious in interpreting the effects of relationship direction; subjects may approach the stimuli in question with strong expectations about the nature of the relationship that will hold. In Seggie's 1975 study, for example, subjects judged whether or not hospitalizing a victim of a tropical disease would improve the chances of recovery. Erlick and Mills' (1967) subjects judged the relationship between the quantity of a particular food a person ate and whether the person felt better or worse. People who believe in the merits of medical science or hearty eating would be likely to expect each to improve general well being. This expectation could produce a bias to report relationships as positive, resulting in errors in judging negatively related and independent events.

Evidence of such an expectation effect was found in the research of Chapman and Chapman (1967 a & b), where subjects judged there to be a positive relationship between semantically-associated clinical signs and symptoms in stimuli that actually presented the sign and symptom as independent, or even negatively related. This illusory correlation effect proved to be highly resistant to a variety of attempts to reduce it, including exposing subjects to the stimuli several times and offering them a $20 reward for accuracy. Similar expectancy effects may be a reason for some past findings of differential accuracy as a function of relationship direction. Any attempt to examine the effect of relationship direction should then be conducted in a context in which prior expectations are minimal.

A second common finding in past research is that judgments of interevent correlations are biased by the relative frequencies of the event states of the variables involved. For example, Jenkins and Ward (1965) asked subjects how much control their responses (pushing Button 1 or 2) had over the frequency with which a score light appeared. Subjects' judgments of control were most strongly correlated with the number of times the score light occurred, regardless of whether that outcome was actually influenced by their choice of buttons. Allan and Jenkins (1980) found that this bias was reduced, but not eliminated when subjects had a single button to press or not to press, compared to Jenkins and Ward's two-button condition (also see Alloy & Abramson, 1979). The findings of these investigations indicate that the probability of the outcome is a second possible confound to be controlled or manipulated in assessing contingency judgment.

A final recurrent finding in past research is that the accuracy of judging interevent contingency depends on how the event frequency information is presented. Two common formats present this information either as a series of

individual event-state combinations (e.g., Alloy & Abramson, 1979; Shaklee & Mims, 1982; Ward & Jenkins, 1965) or as a summary table (e.g., Seggie, 1975; Smedslund, 1963; Ward & Jenkins, 1965). Experiments which have compared the two presentation formats have found accuracy to be higher when the frequency information is summarized in table format.

Of course, the serial and summary formats differ in a variety of ways. Most obvious is the added memory demand involved in the trial-by-trial presentation of information; thus, subjects who add a strong memory load to an already complex judgment process may compromise accuracy to simplify an overwhelming task. Shaklee and Mims (1982) relied upon such a memory account in interpreting their judgment findings. Ward and Jenkins (1965), however, argued that, while important, memory load cannot fully account for the judgment difference between serial and summary formats. Rather, they proposed that the serial presentation of stimulus information may lead subjects to organize the information differently from those who view the same information in a tabled format. In support of this point, Ward and Jenkins note that subjects in their experiments who were shown tabled information after serial presentation used less appropriate judgment strategies than those who saw only the tabled information. If information is organized differently under the two conditions, then this may lead subjects to make different judgments of inter-event relationships. Although this reasoning is plausible, past paradigms have confounded presentation format with memory load; the contributions of memory and organization effects in past research cannot then be separated. The issue is best addressed by comparing use of serial and summary frequency information in conditions alike in memory load.

The present study thus compared serial and summary formats in a setting free of memory demands, while also using a problem for which subjects should

have little bias as to the nature of the interevent relation. The basic situation involved troubleshooting a malfunctioning radio. While this situation is far less dramatic than Polonius' efforts to determine the reason for Hamlet's odd behavior, it is nonetheless representative of everyday instances of causal reasoning.

Subjects were told that an individual was trying to find the cause of an intermittent buzz (B) by occasionally tapping (T) on a wire inside the radio. The results of the troubleshooting were then given to the subject, who was asked to judge the degree to which tapping affected the radio's buzzing: from "causes the sound to occur" to "has no effect on the sound" to "prevents the sound from occurring." This context has the virtue of being one in which subjects should not have a strong expectation about the nature of the response-outcome relationship; tapping a wire should be as likely to complete as to break a loose connection. Similarly, if the wire is not loose, tapping it should have no effect on the buzz.

lding constant the probability of tapping, $p(T)$, both the probability of a buzz given a tap, $p(B/T)$, and the probability of a buzz given no tap, $p(B/\overline{T})$, were systematically varied to yield 24 different troubleshooting conditions. These conditions in turn constituted nine tap-buzz contingencies, $p(B/T) - p(B/\overline{T})$, ranging in .25-steps from -1.00 to +1.00 (see Allan, 1980 for further discussion of various measures of contingency or correlation).

An additional feature of the 24 troubleshooting conditions was that they were contrived in such a way that they varied not only in the tap-buzz contingency, but also in the overall probability per sampling interval of the buzzing sound, $p(B)$. Eight different buzz probabilities were studied, ranging in .125-steps from .125 to 1.000. Because the tap-buzz contingency and the relative frequency of the radio's buzzing vs its not buzzing were independent

dimensions in the present experimental design, the contributions of these variables to subjects' judgments of correlation could be individually assessed.

The method of information presentation was studied with two basic techniques. In one, subjects were given summary tables showing the numbers of times that the four possible event sequences occurred in 24 sampling intervals: tap-buzz, tap-no buzz, no tap-buzz, and no tap-no buzz. In the other, the same information was given in a time line format, with the 24 sampling intervals graphically and linearly arrayed. Such an arrangement preserves the sequential character of the critical events, while minimizing the strong memory demands that are ordinarily placed on subjects when they are given information in a trial-by-trial fashion. This method was originally suggested by Ward and Jenkins (1965, p. 240); however, it has never been utilized in experimental research.

Since past work has not entailed a time line presentation of event frequencies, our series of investigations began by looking at subjects' judgments using this format alone. Experiment 1 explored the effects of tap-buzz contingency and buzz probability on judgments of tap-buzz correlation in both within-subjects and between-subjects paradigms. Experiment 2 directly compared the effects of the time line and summary table methods of information presentation. Because the second experiment disclosed that judgments did differ under the two conditions of information presentation, Experiments 3 and 4 explored possible reasons for the judgment differences.

Experiment 1

The first experiment investigated the judgment of response-outcome correlation when responses and outcomes were shown to subjects in a time line format. In one part of the experiment, each subject received only 1 of 24 possible tap-buzz conditions; in the other part, each subject received all 24

tap-buzz conditions. Both between- and within-subjects conditions were included in order to identify possible influences of multiple judgments, since we hoped to use the more efficient within-subjects procedure in later work. Subjects' ratings of the response-outcome relationships allowed us to determine the degree to which the tap-buzz contingency, $p(B/T) - p(B/\overline{T})$, and the overall probability of the buzzing sound, $p(B)$, influenced their behavior. To determine whether the sign of the response-outcome correlation affected subjects' judgments, equal numbers of positive and negative contingencies were studied.

Method

Subjects. The subjects were participants in an introductory psychology class, who served in the experiment as one option for fulfilling a course requirement. A total of 552 students served in the between-subjects part of the experiment and a total of 25 students served in the within-subjects part.

Problems. A set of 24 problems was constructed. These problems were alike in that they all comprised 24 sampling intervals. Each sampling interval in turn had two components: a "response" component during which a tap might or might not occur, and an "outcome" component during which a buzz might or might not occur. Each of the 48 resulting components of a problem was denoted on the subject's problem sheet as a dash; the 48 consecutive dashes thus constituted the time line for each problem. Taps in the response component of a sampling interval were denoted by an "A" above the dashed time line, and buzzes in the outcome component of a sampling interval were denoted by a "B" below the dashed time line.

For all 24 problems, there were 12 taps represented in the 24 possible response components. Thus, the probability of tapping per sampling interval, $p(T)$, was always .50. Problems varied in terms of the likelihood that a buzz was represented in the outcome components, $p(B)$, and the likelihood of buzzes following taps, $p(B/T)$, and no taps, $p(B/\overline{T})$, in the response components.

For each of the 24 problems, Table 1 shows the numbers of sampling intervals of each of four possible types: tap-buzz, tap-no buzz, no tap-buzz, and no tap-no buzz. Note that the number of sampling intervals with a tap is equal to 12, which is the same as the number of sampling intervals without a tap. Note also that the total number of sampling intervals equals 24. And note finally that the number of sampling intervals with a buzz varies from 3 to 24.

Insert Table 1 about here

For each problem, time lines were constructed from smaller groupings that contained eight sampling intervals. The sequence of event pairings was determined randomly within each eight-sample group. While eight-sampling groups theoretically provide all the necessary information that is needed to distinguish the 24 problems, we thought it advantageous to triple the amount of input given to the subjects in hopes that their judgments might thereby be improved. For example, Problem 18 in Table 1 was represented as follows:

A	A	A	A	A	A	A			A			A	A	A	A
B	B			B		B		B	B	B	B	B		B		B		B	

Figure 1 shows a second method of depicting the 24 problems that were studied. Both the top and bottom portions of the figure locate each problem within the unit square defined by the two independent conditional probabilities, $p(B/T)$ and $p(B/\bar{T})$. The top portion of the figure shows the response-outcome contingency $p(B/T) - p(B/\bar{T})$, of each of the problems; the bottom portion shows the likelihood of the buzzing sound per sampling interval, $p(B)$, for the same problem set. There are nine response-outcome contingencies and eight probabilities of buzz presentation represented by the 24 problems in Figure 1. Furthermore, these two procedural dimensions are orthogonal, as can be seen by

the opposite slopes of the lines that connect the 24 problems in the top and bottom portions of the figure. From the figure it can finally be seen that one possible problem was not included in the set. When $p(B/T) = 0 = p(B/\bar{T})$, $p(B) = 0$; little sense could thus have been made of the task by the subjects (see next section for questionnaire instructions).

Insert Figure 1 about here

Procedure. Subjects were given problem sheets that each contained instructions, a time line, and a rating scale. The instructions read as follows:

> After buying a new radio, Kim finds that it emits a brief buzzing sound every so often. Kim finds this buzzing sound annoying and decides to find its cause. Removing the back of the radio, Kim suspects that a wire may be loose. Kim chooses a wire and taps on it a number of times in order to see if this has any effect on the buzzing sound. In the diagram below, Kim's tapping on the wire is shown by an A above the time line which moves from left-to-right across the page. An occurrence of the brief buzzing sound is shown by a B below the time line.

One of the 24 different time lines then followed. Below the time line was a nine-point rating scale ranging from -4 (prevents sound from occurring) to 0 (has no effect) to +4 (causes sound to occur). Subjects were asked to circle the number that best corresponded to their answer to the question, "If you were Kim, what would you conclude was the effect of tapping on the wire?"

In the between-subjects part of the experiment, only 1 of the 24 problem sheets was given to each subject. In the within-subjects part of the experiment, each subject received all 24 problem sheets, with the order of the sheets randomly determined for each subject. The 24 problem sheets were clipped together; each packet also included the following cover sheet:

> The aim of this experiment is to see how people judge the relationship between their actions and the consequences of those actions. In the 24 sheets that follow, the same basic problem is posed: What is the relation between Kim's tapping on the wire of a malfunctioning radio and the occurrence of a brief buzzing sound that the radio occasionally emits. The 24 sheets differ only in the particular relationship between Kim's tapping and the occurrence of the sound. For each of the 24 sheets, please rate the degree to which Kim's tapping affects the rate of the radio's buzzing, from "prevents the sound from occurring" to "causes the sound to occur." As you go through the 24 problems, you'll soon see that the problems differ from one another to varying degrees. You may sometimes want to look back to prior problems; you may even want to change prior responses. This is OK. It is more important to work through the problems carefully and methodically than to give quick and offhand reactions. Indeed, the materials are paper-clipped together so that you can sort through the many sheets and organize them any way you wish.

Results

Table 2 shows the means and standard deviations of subjects' judgments for the 24 problems in both the between- and within-subjects parts of the experiment. Each of the 24 problems is located in the table by the coordinates $p(B/T) - p(B/\overline{T})$ and $p(B)$. In general, subjects' rating scores were positive functions of both $p(B/T) - p(B/\overline{T})$ and $p(B)$.

Insert Table 2 about here

Figure 2 graphically portrays subjects' rating scores as separate functions of $p(B/T) - p(B/\overline{T})$ and $p(B)$ in each part of the experiment. Analysis of variance simultaneously assessed the reliability of these two sets of functions.

Insert Figure 2 about here

The left panel of Figure 2 displays subjects' ratings as a function of $p(B/T) - p(B/\overline{T})$. The positive diagonal in the figure shows the responses of a

hypothetical judge whose responses correspond in a linear fashion to the actual response-outcome contingencies and who also employs the full rating scale. In the between- and within-subjects parts of the experiment, subjects' judgments were reliable linear functions of $p(B/T) - p(B/\overline{T})$, $F(1, 528) = 139.17$, $p < .001$, and $F(1, 24) = 74.76$, $p < .001$, respectively; however, the slopes of those functions were clearly less than that of our hypothetical linear observer. The between- and within-subjects functions also had reliable quadratic components, $F(1, 528) = 11.28$, $p = .001$, and $F(1, 24) = 28.07$, $p < .001$, respectively; this trend appears to be due to the negative segments of the functions having shallower slopes than the positive segments. Finally, in the within-subjects part of the experiment, the contingency-rating function had a reliable cubic component, $F(1, 24) = 10.96$, $p = .003$; this trend appears to be due to the function having an inverted S shape. Although the overall form of the between-subjects function was similar, it did not have a reliable cubic component.

The right panel of Figure 2 displays subjects' ratings as a function of $p(B)$. In the within-subjects part of the experiment, ratings were a positive linear function of $p(B)$, $F(1, 24) = 32.63$, $p < .001$. In the between-subjects part of the experiment, the linear trend only approached significance, $F(1, 528) = 2.90$, $p = .089$.

To assess the relative contributions of $p(B/T) - p(B/\overline{T})$ and $p(B)$ to subjects' judgment scores, the percentage of problem variance accounted for by these factors was determined through the cubic component of each; beyond the cubic component, no significant variance remained for either part of the experiment. In the between-subjects part of the experiment, $p(B/T) - p(B/\overline{T})$ accounted for 86.47% of the total variance and $p(B)$ accounted for 3.21%; in the within-subjects part of the experiment, the corresponding scores were 71.87% and 24.10%.

Discussion

Subjects' judgmen of contingencies in the time line format showed several interesting trends that were generally comparable in the within- and between-subjects parts of the experime t. These results also accord well with past paradigms using different presentation formats. First, judgments of response-outcome correlation were a reliable function of the contingency between the tapping of a wire and the occurrence of a brief buzzing sound. Subjects' ratings rose as the tap-buzz contingency, $p(B/T) - p(B/\overline{T})$, increased from negative to positive values. Thus, subjects clearly showed some sophistication about appropriate bases of contingency judgment.

The relative accuracy of subjects' judgments is, however, another issue. Mean judgments indicated that subjects rated noncontingent relationships close to zero, but ratings of several negative relationships hovered close to zero as well. While subjects were asked to rate both the degree and the sign of a correlation, the clearest evidence of accuracy here was the rated direction of the relationship. Subjects' judgments should also have been ordered according to the strength of the correlation. While this was generally true, the ratings yielded contingency judgments that were poorer than ideal. Indeed, the quadratic component of the judgment function indicates that subjects did not treat positive and negative relationships symmetrically; contingencies of the same absolute value were rated as stronger for positive than for negative relationships. The form of this difference in ratings of relationship strength closely resembles that found in prior research by Erlick and Mills (1967).

The second main finding was that judgments of correlation were reliably influenced by the likelihood of the buzzing sound, $p(B)$. This bias is comparable to that found in other studies in which the judgment of contingency depended on the likelihood that the outcome occurred (Allan & Jenkins, 1980;

Alloy & Abramson, 1979· Jenkins & Ward, 1965). These prior studies most convincingly demonstrated a bias effect of p(B) with response-outcome contingencies of zero; Allan and Jenkins' (1980) investigation further suggested that the bias effect could arise under positive contingencies. The present report confirms the above trends and also shows that the effect of p(B) on judgments holds under negative response-outcome contingencies as well (see that ratings tend to increase from top to bottom within most columns of Table 2).

Experiment 2

The results of the time line portrayals in Experiment 1 were comparable in many ways to those of past paradigms. However, subjects who view information in a particular format may treat the information in a manner specific to that format; that is, subjects' attention to information may depend on the way the information is presented. The organization or integration of attended information may vary with stimulus format as well. We propose three ways in which the time line and the more familiar summary table format may produce different judgments.

First, tabled presentation of event frequency information offers the subjects tallies of the frequencies of each type of event-state combination. Our time line presentation (like past serial presentation techniques) requires the subjects to generate such tallies on their own. Subjects given time line information may guess rather than count those frequencies, resulting in estimation errors. This logic suggests that judgments with time line presentation will be generally less accurate than judgments with tabled presentation and that such differential accuracy will be relatively constant across positive, negative, and noncontingent relationships. The resultant judgment function should be relatively flat across all contingencies compared to that of tabled information.

A second possible source of difference is the fact that the summary table presents the event-state combinations in a form of comparable salience. In contrast, each type of event pairing has a unique representation in the time line format (i.e., AB, A-, -B, --). As a result, some types of event pairings may be more salient than others. In particular, the interval pairs with two event absences (--) may be less prominent than those with one or both events present. This feature may also have been true of past serial presentation paradigms. If so, subjects should underestimate the frequency of no tap-no buzz pairings. Since the denominator of the conditional probability, $p(B/\overline{T})$, would then be smaller than would be accurate, this would result in an estimate of $p(B/\overline{T})$ that is too high. This in turn should result in a bias to judge contingencies as being more negative in the time line format than the same contingencies presented in the tabled format.

Finally, the time line format allows the subject to determine the delay between tap and buzz that will be counted as a tap-buzz pairing. Consider the interval series A--B. The tabled format would represent this as one occurrence of tap-no buzz and one of no tap-buzz. However, a subject given the time line presentation may well consider this series to be a single pairing of tap-buzz. This tendency would lead to an underestimation of the frequencies of event pairings tap-no buzz and no tap-buzz and an overestimation of the frequency of tap-buzz pairings. These errors would yield an inflated numerator for $p(B/T)$ and a smaller than accurate numerator for $p(B/\overline{T})$. These biases should result in judgments of contingencies being more positive in the time line than in the summary table format. This problem of event segmenting should not have been true of past discrete trial presentations, where each slide or card defined an event-outcome pairing. However, the problem may be true of event processing in real time, when event continua must be defined as discrete events.

Thus, each of three reasons for judgment differences in the two information presentation conditions would result in a unique pattern of judgment outcomes. Whether any of these differences will materialize is an empirical question. Experiment 2 addressed this issue by comparing judgments under the time line format employed in Experiment 1 with judgments of the same problems presented in the summary table format used in past investigations (e.g., Smedslund, 1963; Ward & Jenkins, 1965). Since judgments were so comparable in the between- and within-subjects parts of Experiment 1, subjects in Experiment 2 judged all 24 problems.

Method

Subjects. The subjects were 34 undergraduate research participants.

Problems. The same 24 problems were used here as in Experiment 1. Problems in the time line format were typed on a single sheet of paper with the nine-point rating scale to the right of each problem. Problems in the summary table format were typed on another sheet of paper similar to Table 1, except that the four types of sampling intervals were vertically arrayed; identical rating scales were located beneath each problem. Problems were presented in a single random sequence for the time line format and in a different random sequence for the table format.

Procedure. During the first portion of the experimental session, subjects were given an instruction sheet describing the troubleshooting problems on the attached sheet of paper. For half of the subjects the problems were in the time line format, and for the other half the problems were in the summary table format. During the second half of the session, subjects worked problems in the format not worked in the first half. Instructions for time line problems were the same as those used in Experiment 1. Instructions for summary table problems were the same, with appropriate adjustments to introduce the table rather than the time line format.

Results

Table 3 shows the means and standard deviations of subjects' judgments for the 24 problems given in the time line and summary table formats. Because analysis of variance failed to disclose any reliable effects attributable to the order of format presentation, this factor is not considered in Table 3 nor in later data analysis. As in Experiment 1, subjects' ratings were positive functions of both $p(B/T) - p(B/\bar{T})$ and $p(B)$.

Insert Table 3 about here

Figure 3 graphically depicts subjects' rating scores as separate functions of $p(B/T) - p(B/\bar{T})$ and $p(B)$ for each method of information presentation. Analysis of variance simultaneously compared these two sets of functions.

Insert Figure 3 about here

The left panel of Figure 3 portrays subjects' ratings as a function of $p(B/T) - p(B/\bar{T})$. Overall, ratings were reliable linear, $F(1, 32) = 51.72$, $p < .001$, and quadratic, $F(1, 32) = 12.90$, $p = .001$, functions of tap-buzz contingency. Additionally, there was a reliable quadratic contingency by format interaction, $F(1, 32) = 4.97$, $p = .033$. To pinpoint the source of this interaction, separate analyses of variance were conducted on the time line and summary table data. For both the time line and the summary table formats, ratings were reliable linear functions of contingency, $F(1, 33) = 36.77$, $p < .001$, and $F(1, 33) = 44.27$, $p < .001$, respectively. However, the quadratic trend was reliable for the time line format only, $F(1, 33) = 14.59$, $p = .001$. Thus, subjects' judgments were reliable linear functions of response-

outcome contingency with both methods of information presentation; however, the method of information presentation influenced those functions, with the tabled format supporting judgments that better approximated those of an ideal observer, particularly in the region of negative contingencies.

The right panel of Figure 3 illustrates subjects' ratings as a function of $p(B)$. Overall, ratings were reliable linear, $F(1, 32) = 30.11$, $p < .001$, and quadratic, $F(1, 32) = 26.68$, $p < .001$, functions of outcome probability. Additionally, there were reliable linear, $F(1, 32) = 6.32$, $p = .017$, and quadratic, $F(1, 32) = 12.99$, $p < .001$, outcome probability by format interactions. Because of these interactions, follow-up analyses were separately performed on the time line and summary table data. For the time line data, ratings were reliable linear, $F(1, 33) = 34.57$, $p < .001$, and quadratic, $F(1, 33) = 30.43$, $p < .001$, functions of $p(B)$; for the summary table data, the linear trend was reliable, $F(1, 33) = 5.33$, $p = .027$, and the quadratic trend fell just short of statistical significance, $F(1, 33) = 3.69$, $p = .063$. Thus, the method of information presentation altered the influence of outcome probability on subjects' ratings; providing the information in a time line format both steepened the probability-judgment function and increased its curvature relative to providing the same information in a summary table format.

And, regardless of tap-buzz contingency and buzz probability, judgments were reliably higher in the time line condition than in the summary table condition, $F(1, 32) = 5.03$, $p = .032$.

To assess the relative contributions of response-outcome contingency and outcome probability to subjects' ratings, the percentage of problem variance accounted for by each factor was determined as in Experiment 1. For the summary table data, $p(B/T) - p(B/\overline{T})$ accounted for 81.35% of the total variance and $p(B)$ accounted for 12.58%; for the time line data, the corresponding

scores were 39.48% and 51.79%. Beyond the cubic component, no significant variance remained for the summary table data. For the time line data, the 8.78% remaining variance was small, but statistically significant, $F(17, 561) = 3.23$, $p < .001$.

Discussion

The data from subjects given the time line in this experiment replicate the judgment patterns of subjects in the comparable condition of Experiment 1. In addition, the results of Experiment 2 confirm prior findings (Shaklee & Mims, 1982; Smedslund, 1963; Ward & Jenkins, 1965) that the method of information presentation affects subjects' judgments of response-outcome correlation.

The obtained judgment differences under two conditions comparable in memory demands suggest that past effects of presentation conditions may not be solely attributed to memory. In general, subjects' judgments were more closely attuned to response-outcome contingency when information was given in the summary table than when the same information was given in the time line. First, the contingency-judgment function (left panel of Figure 3) was more symmetrical about zero in the summary table condition, suggesting that subjects rated positive and negative relationships in a comparable fashion. Again, the time line portrayal supported less accurate judgments of negative than positive contingencies. Second, table format judgments were less distorted by the probability of the buzzing sound (right panel of Figure 3). The linear contingency by format interaction showed that the time line judgments were steeper functions of $p(B)$ than the summary table judgments.

We previously reviewed three reasons why time line and summary table formats may result in different contingency judgments. The suggestion that the time line will lead to more errors in estimating frequencies of event

pairings than the summary table predicted overall poorer contingency judgment accuracy (i.e., a flatter, but symmetrical contingency-judgment function) in the time line than in the tabled format condition. The possibility that joint event absences (no tap-no buzz) were less salient in the time line than in the tabled presentation mode predicted a general bias to report relationships as more negative in the time line than in the summary table format. However, neither of these difference patterns describe our results.

Subjects in this experiment did show a tendency to judge relationships as more positive in the time line than in the summary table condition. This result supports our third proposed source of differences, that subjects may group event pairings differently in the time line than the tabled format. In particular, event series A--B could be identified as a single tap-buzz occurrence rather than a tap-no buzz and a no tap-buzz, yielding in a bias to report relationships as positive. However, we should note that while ratings were generally higher in the time line than in the summary table condition, the positivity bias was more pronounced for negative than positive contingencies. One possible account for this finding involves the influence of context on the grouping of event pairings; that is, A--B may be most likely to be judged a tap-buzz occurrence when there are few contiguous AB pairings in the time line, as would be the case in negative contingencies.

Besides helping us to understand why different presentation formats support different judgments, these performance differences between groups also allow us to reject the possibility that time line subjects' problems with rating negative contingencies are due to a response bias or to prior expectations. Any expectation about the effect of tapping on the radio's buzzing should be the same in the two groups, but judgments of negative contingencies were distorted for time line subjects only. Similarly, since subjects made

judgments on the same rating scale in the two conditions, performance differences cannot be attributed to peculiarities in the scale itself.

Experiment 3

The results thus far suggest that subjects may define events differently in the time line and table formats. If this is the principal reason for the inaccurate responses of time line subjects, then their judgments should improve when the continuous stream of events in the time line is separated into discrete units.

Our third experiment further explored the problem of defining individual sampling periods by placing a clear break between paired intervals in the time line format. To do this, we simply added a blank space between successive sampling intervals along the time line. As in the within-subjects part of Experiment 1, subjects rated all 24 tap-buzz contingencies. These judgments were compared to those obtained in Experiment 1, in which successive sampling intervals immediately followed one another.

Method

Subjects. Another group of 25 undergraduate research participants joined the 25 who had served in the within-subjects part of Experiment 1, and whose data are depicted again in the Results section that follows. Subjects in these two groups were from the same introductory psychology course and were tested within 3 weeks of the same school term.

Problems. The problems for the new subjects were identical to those in Experiment 1, except that one blank space was inserted between successive sampling intervals along the time line. This format is illustrated in a sample item (Problem 11):

$$\frac{A}{B}\ \_\ \frac{A}{B}\ \frac{A}{B}\ \_\_\ \frac{A}{B}\ \frac{A}{B}\ \_\ \_\_\ \_\_\ \frac{A}{B}\ \_\ \frac{A}{B}\ \frac{A}{B}\ \_\ \frac{A}{B}\ \_\_\ \_\_\ \frac{A}{B}\ \_\_\ \frac{A}{B}\ \frac{A}{B}\ \_\_\ \_\_\ \frac{A}{B}\ \frac{A}{B}$$

Procedure. The procedure for the new subjects given the broken time lines was identical to that for the former subjects given the unbroken time lines in Experiment 1.

Results

Table 4 shows the means and standard deviations of subjects' judgments for the 24 problems given in the broken and the unbroken time line conditions of Experiment 3. Again, subjects' ratings were positive functions of $p(B/T) - p(B/\overline{T})$ and $p(B)$.

Insert Table 4 about here

Figure 4 graphically illustrates subjects' rating scores as separate functions of $p(B/T) - p(B/\overline{T})$ and $p(B)$ for each time line condition. Analysis of variance simultaneously compared these two sets of functions.

Insert Figure 4 about here

The left panel of Figure 4 shows subjects' ratings as a function of $p(B/T) - p(B/\overline{T})$. Overall, ratings were reliable linear, $F(1, 576) = 542.75$, $p < .001$, quadratic, $F(1, 576) = 34.32$, $p < .001$, and cubic, $F(1, 576) = 20.35$, $p < .001$, functions of tap-buzz contingency. Additionally, there was a reliable linear contingency by time line interaction, $F(1, 576) = 5.08$, $p = .025$, and a near significant quadratic contingency by time line interaction, $F(1, 576) = 3.18$, $p = .075$. Therefore, separate analyses of variance were conducted on the data for the group given the broken time line and for the group given the unbroken time line. For both the broken and unbroken time line groups, ratings were reliable linear functions, $F(1, 24) = 83.74$, $p < .001$,

and $F(1, 24) = 74.76$, $p < .001$, respectively; quadratic functions, $F(1, 24) = 7.17$, $p = .013$, and $F(1, 24) = 28.07$, $p < .001$, respectively; and cubic functions of contingency, $F(1, 24) = 24.83$, $p < .001$, and $F(1, 24) = 10.96$, $p = .003$, respectively. Thus, although the contingency-rating functions were similar, judgments of contingency were more strongly differentiated for subjects in the broken time line group; this greater differentiation was generally more notable for negative than for positive contingencies.

The right panel of Figure 4 portrays subjects' ratings as a function of $p(B)$. Overall, ratings were reliable linear, $F(1, 576) = 139.87$, $p < .001$, and quadratic, $F(1, 576) = 25.33$, $p < .001$, functions of outcome probability. Additionally, there was a reliable quadratic outcome probability by time line interaction, $F(1, 576) = 6.18$, $p = .013$. Separate analyses of variance were therefore conducted on the data from the two time line groups. For both the group given the broken time line and the group given the unbroken time line, ratings were reliable linear functions of $p(B)$, $F(1, 24) = 20.62$, $p < .001$, and $F(1, 24) = 32.63$, $p < .001$, respectively. However, the quadratic trend was reliable for the broken time line group only, $F(1, 24) = 24.01$, $p < .001$. Thus, the probability-rating functions of the two time line groups were similarly sloped, although the function for the broken time line appeared to turn downward at high outcome probabilities more than the function for the unbroken time line.

To assess the relative contributions of response-outcome contingency and outcome probability to subjects' judgments, the percentage of variance accounted for by each factor was determined as in Experiments 1 and 2. For the broken time line group, $p(B/T) - p(B/\overline{T})$ accounted for 77.31% of the total problem variance and $p(B)$ accounted for 19.08%; for the unbroken time line group, the corresponding scores were 71.37% and 24.10%.

Discussion

We introduced the broken time line format in Experiment 3 to partition the time line continuum into discrete sampling intervals. The results of the experiment indicate that this manipulation had an effect on judgments of the problem set. Subjects judging broken time lines showed greater differentiation in their ratings as a function of the scheduled contingency than subjects judging unbroken time lines. This increased differentiation was generally more prominent for negative than for positive relationships, a difference which was also true of subjects judging tabled information in Experiment 2.

Thus, the results of subjects who viewed the broken time lines duplicate in some respects the behavior of subjects judging on the basis of tabled information. Our ability to increase the accuracy of contingency judgments by this manipulation enhances confidence in our interpretation that subjects made errors in identifying discrete event pairings in the continuous time lines. The similarity of judgments of tabled and broken time line information suggests that one function of the table may be to separate a stream of events into coherent units. Such units may be more readily classed according to the type of event pairing and thus may be more accurately incorporated into a contingency judgment.

While breaking the flow of the time line into discrete sampling intervals yielded judgments more similar to those made with summary table presentation, inspection of Figures 3 and 4 shows that the judgments obtained under these two conditions were not identical. Contingency-judgment functions under the broken time line format were less symmetrical about zero than under the summary table format, and probability-rating functions were steeper in the former condition than in the latter. Thus, other factors may well contribute to the differences in contingency judgments obtained with the time line and summary table formats in Experiment 2.

Experiment 4

Thus far, our leading interpretation of the problems created by a continuous representation of events is that people have difficulty breaking the stream into discrete units. An alternative approach to testing this account might be to teach people to parse the time line into the component units. If such training produces judgment functions like those found in our broken time line and table formats, such findings would further support this as the source of judgment differences. A second function of the table mentioned earlier might be to offer subjects numerical summaries of the information about the four event combinations. This summarized information may be more readily incorporated into a decision rule in judging event covariations. In this way, judgment accuracy might be further enhanced if subjects were asked to count the occurrences of each event-state combination and note these frequencies in a table. By this process, subjects would effectively convert a time line into a table format.

Our fourth and final experiment used each of these approaches. One group of subjects was presented with the 24 problems in our original time line format, but were taught to break the line into response-outcome intervals (line-interval). A second group received these instructions and were also asked to count the frequencies of each event-state pairing and write those frequencies in a table (line-table). Time line and table groups using our original instructions served as comparison conditions for these manipulations. Improved judgment by line-interval subjects compared to time line subjects would further implicate line segmenting as a factor in contingency judgment. Further improvements by line-table subjects would suggest that summary information is also an important function of the tabled format. Because we found sex differences in contingency judgment in related work of ours (Shaklee & Hall, in press), sex was included as a factor in this experiment.

Method

Subjects. A total of 160 introductory psychology subjects served in the experiment with 20 males and 20 females in each of four judgment conditions.

Problems. The 24 contingencies for this experiment were the same as those in the previous experiments. Format of problems in the time line and table representations was the same as that used in Experiments 1 and 2.

Procedure. The introduction to the troubleshooting problems was identical to that used in the previous studies, except that the problem representation was explained in one of four ways:

Line: These instructions were the same as those used in Experiments 1 and 2.

Line-Interval: These problems were represented in a time line like that used in Experments 1 and 2, but in this case subjects were specifically instructed how to break the time line into response-outcome intervals. Instructions were as follows:

> Each dash on the time line represents one unit of time. Time units come in pairs, with the first an opportunity for a response (Tap or No Tap) and the second an opportunity for an outcome (Buzz or No Buzz). Thus, pairs of successive intervals can be of four types: Tap-Buzz, Tap-No Buzz, No Tap-Buzz, No Tap-No Buzz. For each of the time lines, please rate the degree to which Kim's tapping affects the rate of the radio's buzzing, from "prevents the sound from occurring" to "causes the sound to occur."

Line-Table: Problems and instructions were identical to those in the Line-Interval condition, except that each problem was accompanied by a blank table labeled as in the previous table condition of Experiment 2. Subjects were instructed to complete the table before making their judgment. Instructions were as follows:

> Each dash on the time line represents one unit of time. Time units come in pairs, with the first an opportunity for a response (Tap or No Tap) and the second an opportunity for

> an outcome (Buzz or No Buzz). Thus, pairs of successive intervals can be of four types: Tap-Buzz, Tap-No Buzz, No Tap-Buzz, No Tap-No Buzz. For each time line, please count the frequency of each of these four types of interval pairs. Enter those frequencies in the table to the right of the time line. Once you have completed the table, please rate the degree to which Kim's tapping affects the rate of the radio's buzzing, from "prevents the sound from occurring" to "causes the sound to occur."

Table: Problems and instructions in this condition were identical to those in Experiment 2.

In each condition, the information offered in the instructions was shown on a sample problem illustrating each type of response-outcome pairing. Subjects were invited to ask any questions they might have, after which they proceeded at their own pace through the problem set.

Results

Means and standard deviations of subjects' judgments for the 24 problems in each judgment condition are shown in Table 5. Figure 5 illustrates subjects' judgments of the nine contingencies, $p(B/T) - p(B/\overline{T})$, and the eight probabilities of buzzing sound, $p(B)$, for the four judgment conditions. These functions were simultaneously compared by analysis of variance, including sex of subject and judgment condition as factors. Paired follow-up analyses were conducted on interactions, setting alpha at .025 to reduce the experiment-wide error rate.

Insert Table 5 and Figure 5 about here

The overall analysis yielded reliable linear, $F(1, 152) = 851.86$, $p < .001$. quadratic $F(1, 152) = 100.92$, $p < .001$, and cubic $F(1, 152) = 12.52$, $p < .001$ trends of response-outcome contingency on subjects' judgments. As in our previous experiments, ju ments were a function of problem contingency, but

with judgments of negative relations closer to zero than those of positive relations. This analysis also showed a main effect of judgment condition, $F(3, 152) = 11.40$, $p < .001$, although that effect is qualified by a contingency by condition interaction, $F(23, 3496) = 2.47$, $p < .001$. As seen in the left portion of Figure 5, the form of this interaction shows that judgments in the Table condition were most symmetrical about zero, judgments in the Line condition were least symmetrical, and judgments in the Line-Interval and Line-Table conditions fell between these two extremes. Follow-up analyses compared contingency judgment functions for selected condition pairs. Line-Interval and Line conditions were compared to identify the effect of the interval segmenting instructions. This analysis showed Line-Interval subjects to be significantly different from Line subjects; linear trend $F(1, 76) = 11.12$, $p = .001$, the quadratic trend approaching significance $F(1, 76) = 4.92$, $p = .029$. Comparison of Line-Table and Line-Interval contingency functions showed that tabling the frequency information had no additional effect on judgment accuracy. Line-Table and Table judges were compared to see if judges who tabled the frequency information for themselves were equivalent in judgment to those who judged tables provided by the experimenter. This comparison showed that contingency judgment functions were not equivalent for the two groups, with Line-Table and Table judges reliably different in quadratic trend, $F(1, 76) = 5.83$, $p = .018$, but not in linear or cubic trends.

Sex differences in contingency functions were statistically significant, with the contingency-judgment function for females flatter than that for males: linear trend $F(1, 152) = 3.94$, $p = .049$, cubic trend $F(1, 152) = 4.38$, $p = .038$. This sex effect did not interact significantly with judgment condition.

As in our previous experiments, subjects' judgments were an increasing function of the probability of the buzzing sound (see right portion of Figure

5). Ratings showed significant linear, $F(1, 152) = 210.66$, $p < .001$, quadratic, $F(1, 152) = 80.90$, $p < .001$, and cubic, $F(1, 152) = 4.58$, $p = .034$, trends as a function of $p(B)$. Unlike previous analyses, however, these probability-judgment functions were not reliably affected by judgment condition, although the Line group again showed the greatest effect of $p(B)$ and the Table group showed the least effect. Effects of $p(B)$ also did not differ as a function of subjects' sex.

The relative contributions of response-outcome contingency and outcome probability in each of the four conditions were determined as in the prior experiments. For the Table group, $p(B/T) - p(B/\overline{T})$ accounted for 89.07% of the total problem variance and $p(B)$ accounted for 9.47%; for the Line-Table group, the corresponding scores were 80.97% and 17.02%; for the Line-Interval group, the scores were 76.04% and 17.61%; and for the Line group, the scores were 71.38% and 22.64%. In only the latter two groups was the residual variance significant: Line-Interval residual = 6.35%, $F(17, 646) = 6.72$, $p < .001$, and Line residual = 5.98%, $F(17, 646) = 2.25$, $p = .003$.

Since frequency judgment errors may detract from contingency judgment accuracy, the frequency tables generated by subjects in the Line-Table condition were examined for accuracy. Overall, errors were small, with mean absolute deviations of .15, .10, .30, and 1.65 for Tap-Buzz, Tap-No Buzz, No Tap-Buzz, and No Tap-No Buzz frequencies, respectively. In view of the differential judgments of positive and negative relationships in this condition, frequency judgment accuracy was compared for problems representing positive and negative contingencies. Absolute deviations were averaged across table cells for this analysis. A matched-pairs t-test showed no reliable differences in frequency judgment errors on positive and negative contingencies, $t(39) < 1$.

Discussion

Experiment 4 represents a conceptual replication of our third experiment.

In Experiment 3, we broke the time line into discrete units. In this experiment, we taught the subjects themselves to define these intervals. The results indicate that the manipulations in the two experiments had similar effects. Line-Interval and Line-Table subjects in Experiment 4 produced contingency-judgment functions intermediate to those of our Line and Table subjects. Line-Interval and Line-Table subjects' contingency-judgment functions were more symmetrical about zero than that of Line subjects, although the two new conditions did not differ from each other. This failure to find additional improvement by subjects who completed a frequency table indicates that the availability of summary information contributes little to judgment accuracy. However, the similarity of these two functions to that of subjects in our past broken time line condition enhances our confidence in the problem of event segmenting as a source of error in judging negative relationships.

The finding that Line-Table judges are also less accurate than Table judges is a bit of a surprise. These subjects have effectively converted time line information into a tabled format. However, the accuracy of that conversion is a second question. Since any deviations in frequency judgments must necessarily be in the direction of lower accuracy, subjects in this condition may have somewhat erroneous information on which to base their judgments. However, a look at subjects' frequency counts indicates reasonable accuracy; indeed, 12 out of 40 subjects did not show a single error on any of the 24 problems. In addition, error rates were similar on negative and positive contingency problems. Thus, inaccuracy of frequency judgments constitutes a weak account of the difference in judgment functions of Line-Table and Table subjects.

These differences between Line-Table and Table judgments replicate the stimulus presentation effects of Ward and Jenkins (1965) in a substantially

different format. Their subjects viewed sequences of event-outcome pairs (cloud seeding or not/rain or no rain), each sequence indicating some degree of positive relationship. When the sequence was complete, one group of subjects saw a table summarizing the frequencies of each of the event-state combinations. A second group saw the tabled information only. Ward and Jenkins found that subjects who saw the tabled information after the event series were less accurate in their judgments than those who saw the tabled information alone. It was this finding that inspired the experimenters to conclude that viewing the event sequence had caused the subjects to represent the information in a way that the table failed to counteract, perhaps differentially emphasizing the relative importance of particular event-state pairings. Our own results parallel these past findings closely. In our case, however, subjects viewed event contingencies in a linear representation free of memory demands.

As in our previous experiments, subjects' judgments here were biased by the probability of the buzzing sound. However, unlike Experiments 2 and 3, the extent of that bias was not reliably different in the Line and Table judgment conditions. The failure to replicate this finding is surprising and difficult to account for given the comparability of other aspects of the present results to our other previous outcomes. This finding does temper our confidence in the previous result that judgments of tabled information are relatively free of the effect of the probability of outcome.

Finally, this experiment showed a reliable effect of sex, with contingency-judgment functions of females reliably flatter than those of males. This difference may indicate that females have a higher judgment error rate than males, contributing to flatter functions. This interpretation is congruent with findings in our related work (Shaklee & Hall, in press) showing

that females use simpler, less accurate rules than those used by males to judge event covariations. An alternative interpretation of the sex differences in the present experiment is that the two sexes judge the problems with similar accuracy, but that the females use a more limited range of the scale to make their judgments. However, a comparison of judgments indicates that the two sexes use the scale extremes (±4) at comparable rates (11.3% and 12.2% of judgments for males and females, respectively), ruling out response conservatism as a viable account of this sex difference.

Concluding Comments

In overview, the results of four different experiments suggest that judgments of interevent contingency importantly depend on the method of presenting information about event pairings. Most accurate were judgments of summary table information (Experiments 2 and 4); least accurate were judgments of information presented in a continuous time line format (Experiments 1, 2, and 4). The accuracy of subjects judging partitioned time lines (Experiment 3) fell in between that of the other two conditions. Subjects trained to segment continuous time lines (Experiment 4) made judgments similar to those who saw partitioned time lines. This evidence suggests that Ward and Jenkins (1965) were correct in their suspicion that presentation format may influence subjects' treatment of frequency information in making contingency judgments. Our evidence indicates that subjects may break event sequences into different discrete event pairings depending on the format in which the frequency information is presented. This explanation accounts well for our own findings, but may not be similarly useful in explaining the effects of relationship direction in some past paradigms. As noted earlier, slide or card sequence presentations offer event pairings as discrete units rather than as event continua.

This interpretation offers a ready account for the finding in past research that subjects judge negative relationships less accurately than

positive relationships. Past researchers have suggested that subjects know how to judge positive, but not negative contingencies. Allan (1980), however, pointed out one difficulty with this interpretation; subjects who only know how to judge positive relationships must be able to distinguish between positive and negative relationships in order to apply the appropriate rule to positive contingencies. Presumably, a different, less accurate rule is applied to negative (and independent) relationships. Thus, this interpretation requires that an individual maintain more than one rule to judge event contingencies, and that the person know when to apply which rule to which relationship.

Our analysis indicates a single judgment problem which would result in differential accuracy on positive and negative relationships: that is, subjects' boundaries for event segments depend on the other events in the stream. Positive relationships are typified by many response-outcome pairs which would define a brief time interval as a response-outcome unit. However, where few outcomes promptly follow responses, the observer may accept relatively delayed outcomes as "caused" by the response. The estimate of response-outcome pairs is inflated, resulting in an illusion of a relationship which is less negative than is objectively the case.

We would argue that the problems our subjects encountered in the time line format could be similar to those encountered in judgments of real world contingencies--response-outcome delays may vary in everyday experience. One task of the perceiver is then to define which sequences represent true response-outcome pairings. Investigations of the cues used to break event sequences into discrete units are rare. Our evidence suggests that understanding this process may be important to our ability to account for contingency judgments.

References

Allan, L. A note on measurements of contingency between two binary variables in judgment tasks. Bulletin of the Psychonomic Society. 1980, 415, 147-149.

Allan, L., & Jenkins, H. The judgment of contingency and the nature of the response. Canadian Journal of Psychology/Review of Canadian Psychology, 1980, 34, 1-11.

Alloy, L., & Abramson, L. Judgment of contingency in depressed and nondepressed students: Sadder but wiser? Journal of Experimental Psychology: General, 1979, 108, 441-485.

Chapman, L., & Chapman, J. Genesis of popular but erroneous psychodiagnostic observations. Journal of Abnormal Psychology, 1967, 72, 193-204. (a)

Chapman, L., & Chapman, J. Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. Journal of Abnormal Psychology, 1967, 74, 271-280. (b)

Crocker, J. Judgment of covariation by social perceivers. Psychological Bulletin, 1981, 90, 272-292.

Erlick, D. E., & Mills, R. G. Perceptual quantification of conditional dependency. Journal of Experimental Psychology, 1967, 43, 9-14.

Heider, F. The psychology of interpersonal relations. New York: Wiley, 1958.

Hume, D. A treatise of human nature. In A. Flew (Ed.), On human nature and the understanding. New York: Collier, 1962. (A Treatise of human nature was originally published in 1739.)

Inhelder, B., & Piaget, J. The growth of logical thinking from childhood to adolescence. New York: Basic Books, 1958.

Jenkins, H., & Ward, W. Judgment of contingency between responses and outcomes. Psychological Monographs, 1965, 79, 1-17.

Kelley, H. H. Attribution theory in social psychology. In D. Levine (Ed.), Nebraska Symposium on Motivation (vol. 15). Lincoln: University of Nebraska Press, 1967.

Mackintosh, N. J. The psychology of animal learning. London: Academic Press, 1974.

Morgan, C. L. The limits of animal intelligence. Fortnightly Review, 1893. 54, 223-239.

Morgan, C. L. An introduction to comparative psychology. London: Walter Scott, Ltd., 1894.

Rescorla, R. A. Some inplications of a cognitive perspective on Pavlovian conditioning. In S. H. Hulse, H. Fowler, & W. K. Honig (Eds.), Cognitive processes in animal behavior. Hillsdale, NJ: Erlbaum, 1978.

Seggie, J. The empirical observation of the Piagetian concept of correlation. Canadian Journal of Psychology/Review of Canadian Psychology, 1975, 29, 32-42.

Seggie, J., & Endersby, H. The empirical implications of Piaget's concept of correlation. Australian Journal of Psychology, 1972, 24, 3-8.

Shaklee, H., & Hall, L. Methods of assessing strategies for judging covariation between events. Journal of Educational Psychology, in press.

Shaklee, H., & Mims, M. Sources of error in judging event covariations: Effects of memory demands. Journal of Experimental Psychology: Learning, Memory, and Cognition, 1982, 8, 208-224.

Shaklee, H., & Tucker, D. A rule analysis of judgments of covariation between events. Memory and Cognition, 1980, 8, 459-467.

Smedslurd, J. The concept of correlation in adults. Scandinavian Journal of Psychology, 1963, 4, 165-173.

Ward, W., & Jenkins. H. The display of information and the judgment of contingency. Canadian Journal of Psychology, 1965, 19, 231-241.

Footnote

This research was supported by NSF grant 79-14160 to E.A.W. and NIE grant G-80-0091 to H.S. The authors are greatly indebted to R. H. Hohle for his helpful technical assistance. Portions of this research were reported at the annual meeting of the Psychonomic Society, Philadelphia, PA, November, 1981. Requests for reprints may be sent to either author, Department of Psychology, The University of Iowa, Iowa City, IA 52242.

Table 1

Frequencies of Response-Outcome Possibilities in Each Experimental Problem

Problem	Tap-Buzz	Tap-No Buzz	No Tap-Buzz	No Tap-No Buzz
1	12	0	0	12
2	9	3	0	12
3	6	6	0	12
4	3	9	0	12
5	12	0	3	9
6	9	3	3	9
7	6	6	3	9
8	3	9	3	9
9	0	12	3	9
10	12	0	6	6
11	9	3	6	6
12	6	6	6	6
13	3	9	6	6
14	0	12	6	6
15	12	0	9	3
16	9	3	9	3
17	6	6	9	3
18	3	9	9	3
19	0	12	9	3
20	12	0	12	0
21	9	3	12	0
22	6	6	12	0
23	3	9	12	0
24	0	12	12	0

Table 2

Means and Standard Deviations (in Parentheses) of Subjects' Ratings in the Between- and Within-Subjects Parts of Experiment 1

	$p(B/T) - p(B/\bar{T})$								
p(B)	-1.00	-0.75	-0.50	-0.25	0.00	+0.25	+0.50	+0.75	+1.00
				Between Subjects					
.125				-1.57 (1.53)		0.13 (0.90)			
.250			-0.91 (1.59)		0.09 (1.84)		1.30 (1.57)		
.375		-1.04 (2.07)		-0.74 (1.48)		0.17 (1.79)		1.61 (1.52)	
.500	-1.43 (2.10)		0.00 (1.87)		-0.13 (2.05)		0.96 (1.49)		2.30 (2.37)
.625		-0.39 (1.83)		-0.52 (2.00)		0.39 (1.69)		1.78 (1.69)	
.750			-0.30 (1.97)		0.00 (1.14)		1.61 (1.44)		
.875				-0.52 (2.02)		0.09 (2.02)			
1.000					0.09 (0.88)				
				Within Subjects					
.125				-1.48 (1.36)		-0.52 (1.65)			
.250			-0.60 (1.94)		-0.60 (1.72)		0.88 (1.63)		
.375		-0.92 (1.85)		-0.48 (1.10)		0.40 (0.94)		1.96 (1.31)	
.500	-1.16 (1.78)		0.00 (1.36)		0.08 (1.49)		1.52 (1.45)		3.48 (1.42)
.625		0.20 (1.39)		0.12 (1.27)		1.28 (1.22)		2.24 (1.24)	
.750			0.44 (1.39)		0.60 (1.20)		2.12 (1.11)		
.875				1.28 (1.46)		1.48 (1.58)			
1.000					0.92 (1.90)				

Table 3

Means and Standard Deviations (in Parentheses) of Subjects' Ratings Under the Time Line and Summary Table Formats of Experiment 2

	$p(B/T) - p(B/\bar{T})$								
p(B)	-1.00	-0.75	-0.50	-0.25	0.00	+0.25	+0.50	+0.75	+1.00
Time Line									
.125				-2.38 (2.06)		-2.09 (2.11)			
.250			-1.09 (2.65)		-1.15 (2.20)		0.56 (2.19)		
.375		-1.32 (1.81)		-0.62 (1.78)		0.94 (1.24)		1.41 (1.97)	
.500	-0.94 (2.11)		-0.26 (1.38)		-0.06 (1.75)		1.29 (1.74)		2.47 (2.29)
.625		0.62 (1.85)		0.32 (1.34)		1.29 (1.15)		1.85 (1.80)	
.750			0.71 (1.72)		0.85 (1.54)		1.85 (1.77)		
.875				1.76 (2.04)		1.62 (2.00)			
1.000					0.79 (2.26)				
Summary Table									
.125				-1.41 (2.18)		-0.21 (2.18)			
.250			-1.09 (2.72)		-0.38 (1.91)		0.74 (2.36)		
.375		-1.03 (2.55)		-1.03 (2.02)		0.29 (1.49)		1.26 (1.82)	
.500	-1.44 (2.17)		-0.74 (1.87)		0.24 (1.06)		1.15 (1.65)		2.44 (1.87)
.625		-1.68 (1.74)		-0.06 (1.55)		1.03 (1.54)		1.24 (2.38)	
.750			-0.29 (2.01)		0.50 (1.54)		1.62 (1.97)		
.875				0.38 (2.26)		0.91 (2.12)			
1.000					0.50 (1.74)				

Table 4

Means and Standard Deviations (in Parentheses) of Subjects' Ratings Under the Broken and Unbroken Time Line Conditions of Experiment 3

	$p(B/T) - p(B/\bar{T})$								
p(B)	-1.00	-0.75	-0.50	-0.25	0.00	+0.25	+0.50	+0.75	+1.00
				Broken Time Line					
.125				-1.64 (1.98)		-0.48 (1.98)			
.250			-1.36 (1.62)		-0.28 (1.37)		0.96 (1.56)		
.375		-0.96 (2.09)		0.36 (1.16)		0.56 (1.17)		2.24 (1.77)	
.500	-2.12 (2.63)		0.16 (0.83)		0.36 (1.32)		1.60 (1.36)		3.80 (0.98)
.625		-0.60 (1.96)		0.52 (1.02)		1.68 (1.26)		2.08 (1.57)	
.750			0.12 (1.39)		1.12 (1.34)		1.92 (1.44)		
.875				0.84 (1.41)		1.52 (1.47)			
1.000					0.24 (0.86)				
				Unbroken Time Line					
.125				-1.48 (1.36)		-0.52 (1.65)			
.250			-0.60 (1.94)		-0.60 (1.72)		0.88 (1.63)		
.375		-0.92 (1.85)		-0.48 (1.10)		0.40 (0.94)		1.96 (1.31)	
.500	-1.16 (1.78)		0.00 (1.36)		0.08 (1.49)		1.52 (1.45)		3.48 (1.42)
.625		0.20 (1.39)		0.12 (1.27)		1.28 (1.22)		2.24 (1.24)	
.750			0.44 (1.39)		0.60 (1.20)		2.12 (1.11)		
.875				1.28 (1.46)		1.48 (1.58)			
1.000					0.92 (1.90)				

Table 5

Means and Standard Deviations (in Parentheses) of Subjects' Ratings in the Four Conditions of Experiment 4

	$p(B/T) - p(B/\bar{T})$								
p(B)	-1.00	-0.75	-0.50	-0.25	0.00	+0.25	+0.50	+0.75	+1.00
				Line					
.125				-1.93 (1.97)		-0.78 (2.27)			
.250			-0.78 (1.84)		-0.55 (1.72)		1.15 (1.77)		
.375		-0.98 (1.93)		-0.45 (1.73)		0.70 (1.99)		2.13 (1.60)	
.500	-1.28 (1.95)		-0.25 (1.32)		0.05 (1.53)		1.58 (1.53)		3.45 (1.16)
.625		0.45 (1.84)		0.25 (1.32)		1.23 (1.33)		2.25 (1.32)	
.750			0.55 (1.99)		0.55 (1.72)		1.60 (1.77)		
.875				0.60 (2.30)		1.83 (1.66)			
1.000					0.68 (2.08)				
				Line-Interval					
.125				-2.33 (1.52)		-0.33 (2.04)			
.250			-2.10 (1.69)		-0.58 (1.46)		1.48 (1.38)		
.375		-1.80 (1.93)		-0.60 (1.26)		1.28 (0.89)		2.60 (1.02)	
.500	-2.55 (1.48)		-0.80 (1.31)		0.63 (1.07)		1.70 (0.95)		3.80 (0.64)
.625		-0.15 (1.73)		0.23 (1.15)		1.15 (1.11)		2.70 (0.81)	
.750			0.63 (1.20)		0.63 (1.09)		2.13 (1.05)		
.875				0.80 (1.68)		2.08 (1.49)			
1.000					0.20 (1.42)				

Table 5 (continued)

p(B)	p(B/T) - p(B/$\bar{T}$)								
	-1.00	-0.75	-0.50	-0.25	0.00	+0.25	+0.50	+0.75	+1.00
Line-Table									
.125				-1.90 (1.39)		-0.70 (1.91)			
.250			-1.60 (2.31)		-0.63 (1.43)		0.58 (1.50)		
.375		-2.48 (1.60)		-1.30 (1.27)		0.50 (1.10)		2.20 (1.49)	
.500	-2.28 (1.79)		-0.88 (1.35)		0.20 (0.90)		1.63 (1.70)		3.68 (0.98)
.625		-0.73 (1.57)		0.08 (1.23)		1.23 (1.21)		2.70 (1.38)	
.750			-0.05 (1.52)		0.43 (1.28)		2.08 (1.44)		
.875				0.33 (1.54)		1.68 (1.47)			
1.000					0.20 (1.40)				
Table									
.125				-2.03 (1.42)		-0.25 (1.76)			
.250			-1.90 (1.76)		-0.38 (1.35)		0.68 (1.79)		
.375		-2.20 (1.44)		-1.20 (1.31)		0.53 (1.40)		1.93 (1.99)	
.500	-3.00 (1.67)		-1.73 (1.28)		-0.03 (0.47)		1.65 (1.35)		3.10 (1.77)
.625		-1.83 (1.72)		-0.70 (1.10)		1.20 (1.03)		2.78 (0.88)	
.750			-0.53 (1.76)		0.58 (0.92)		1.98 (1.35)		
.875				-0.43 (1.56)		1.58 (1.20)			
1.000					0.35 (1.26)				

Figure Captions

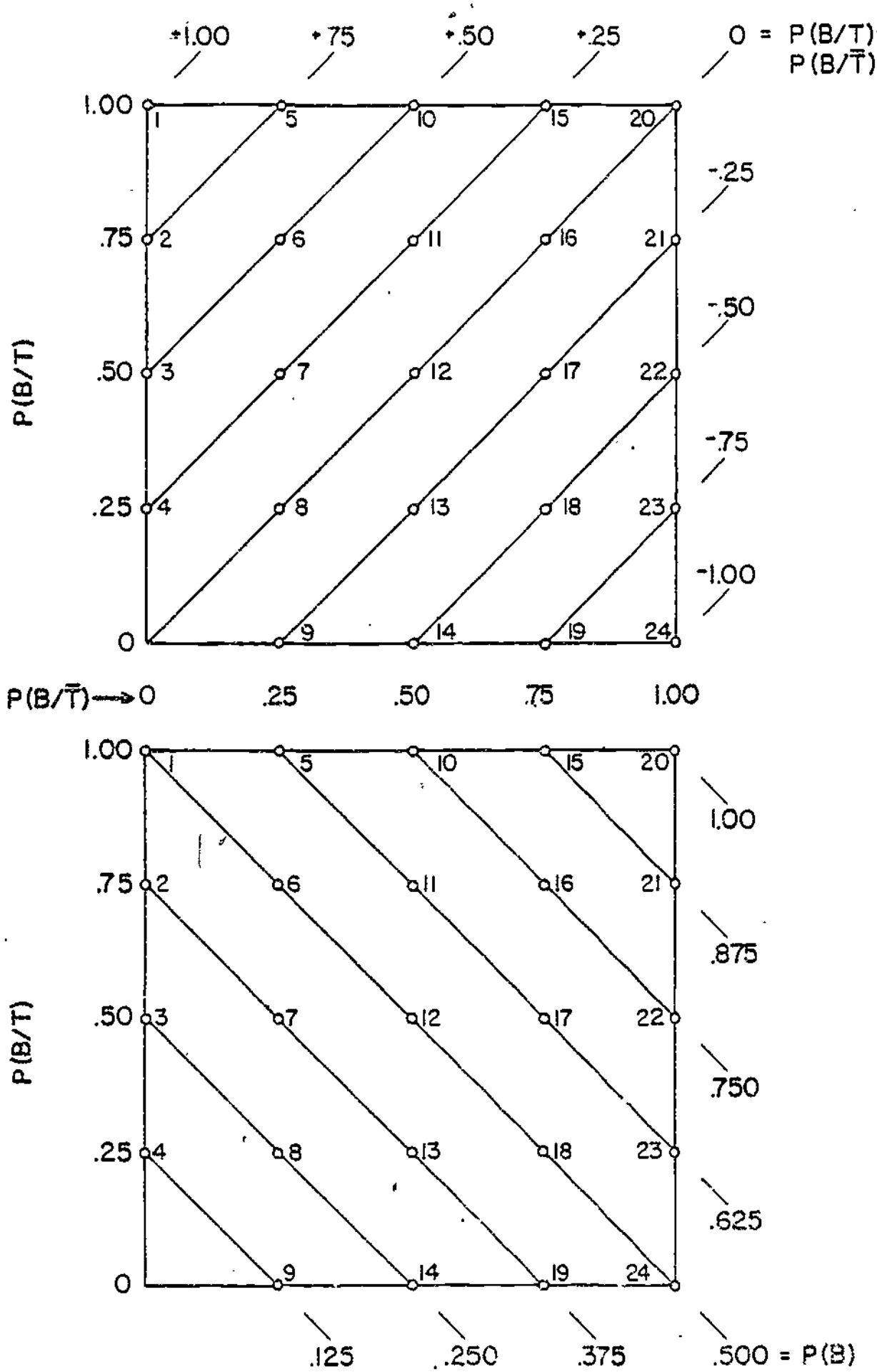
Figure 1. The 24 different response-outcome problems on the coordinates $p(B/T)$ and $p(B/\overline{T})$. The top portion locates the nine different response-outcome contingencies, $p(B/T) - p(B/\overline{T})$, on the unit square; the bottom portion locates the eight different outcome probabilities, $p(B)$. See text for additional explanation.

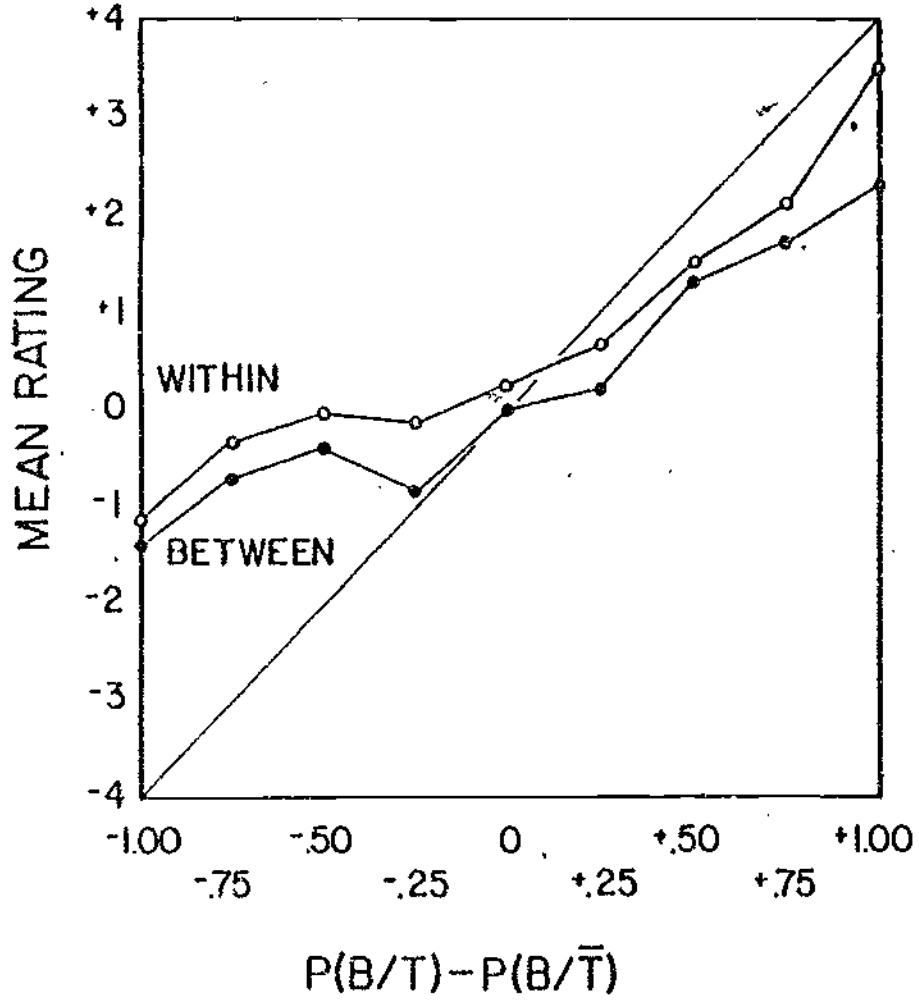
Figure 2. Contingency-judgment functions (left) and probability-judgment functions (right) in the within- and between-subjects parts of Experiment 1.

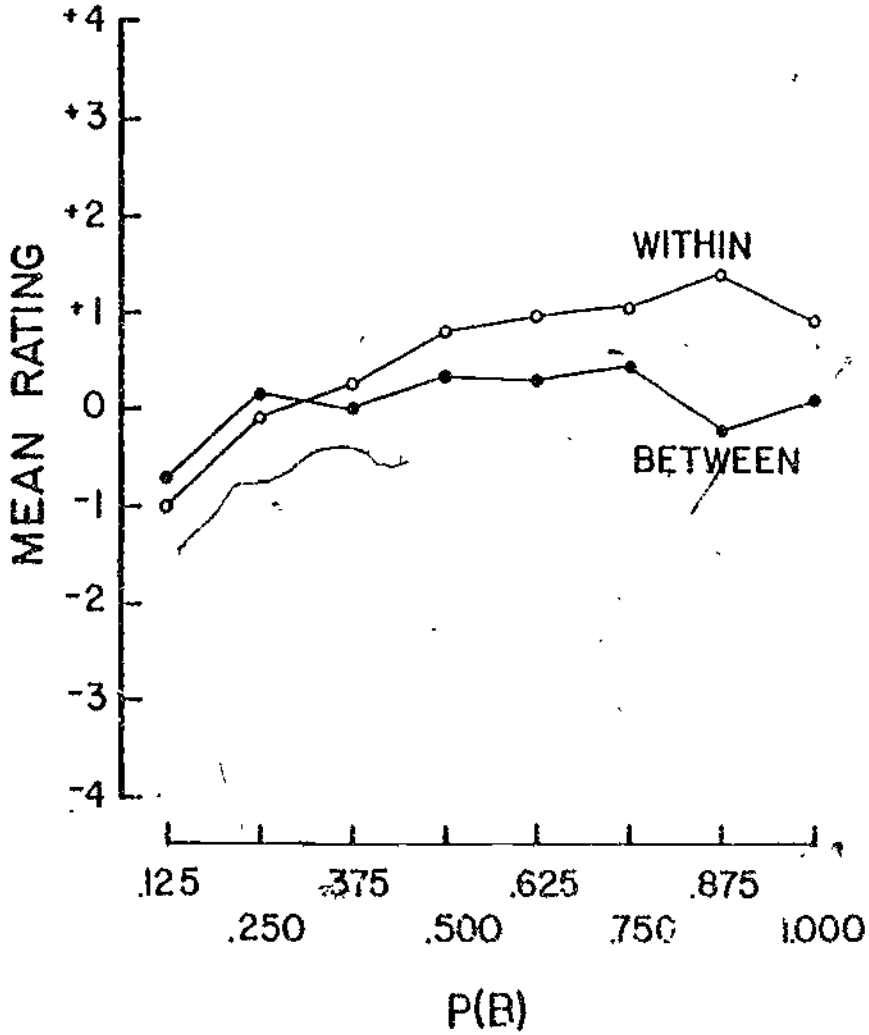
Figure 3. Contingency-judgment functions (left) and probability-judgment functions (right) under the time line and summary table formats of Experiment 2.

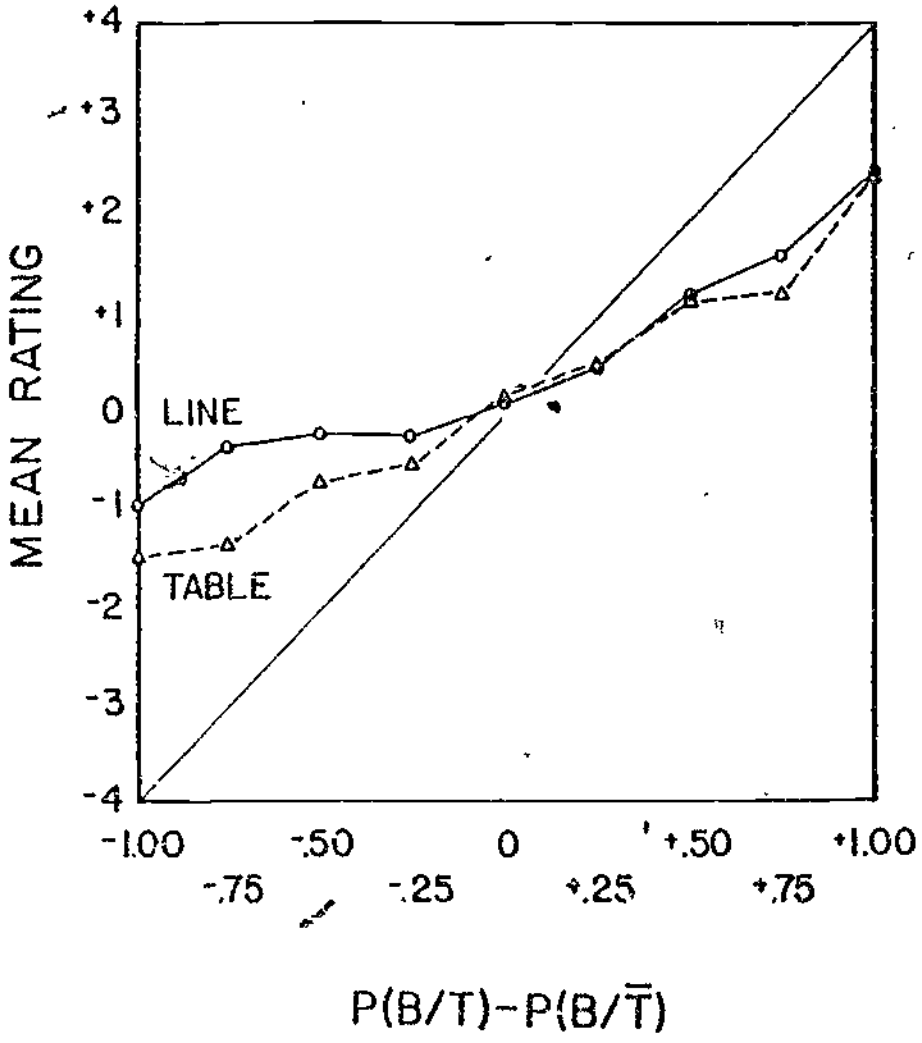
Figure 4. Contingency-judgment functions (left) and probability-judgment functions (right) under the broken and unbroken time line conditions of Experiment 3.

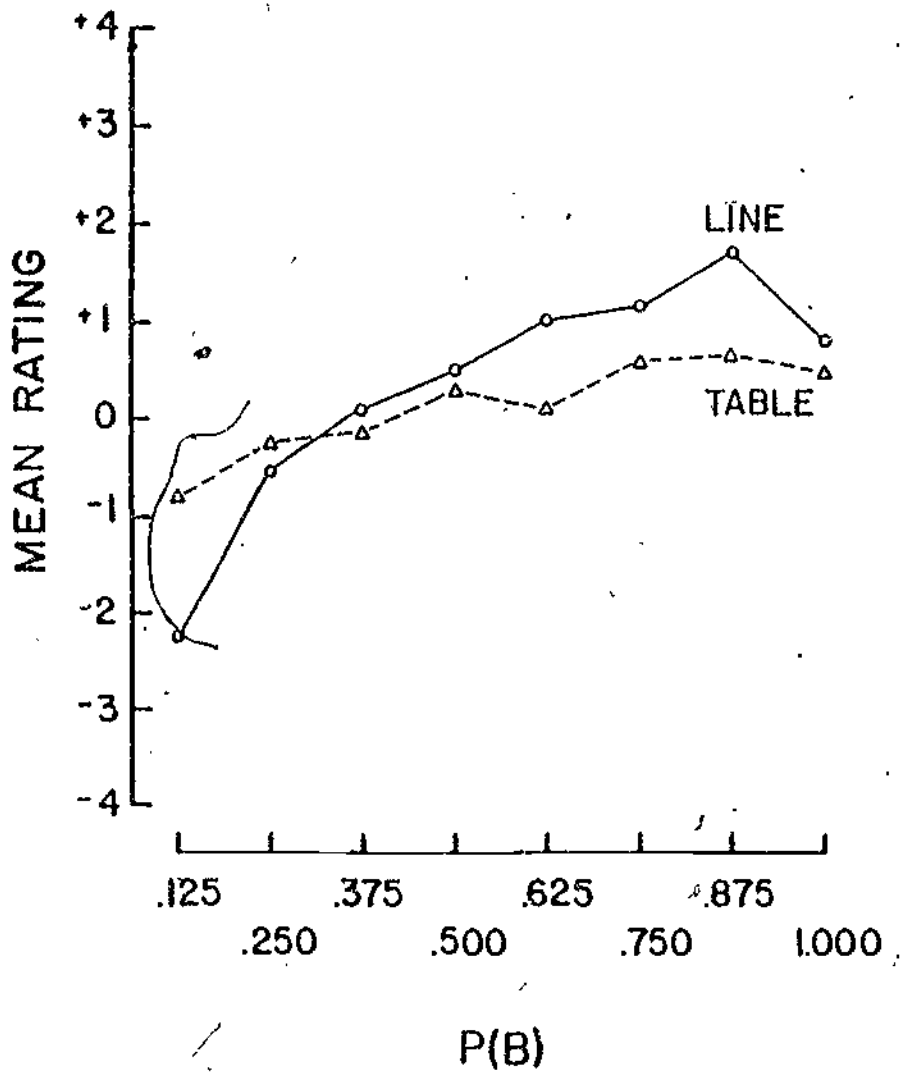
Figure 5. Contingency-judgment functions (left) and probability-judgment functions (right) under the four experimental conditions of Experiment 4.

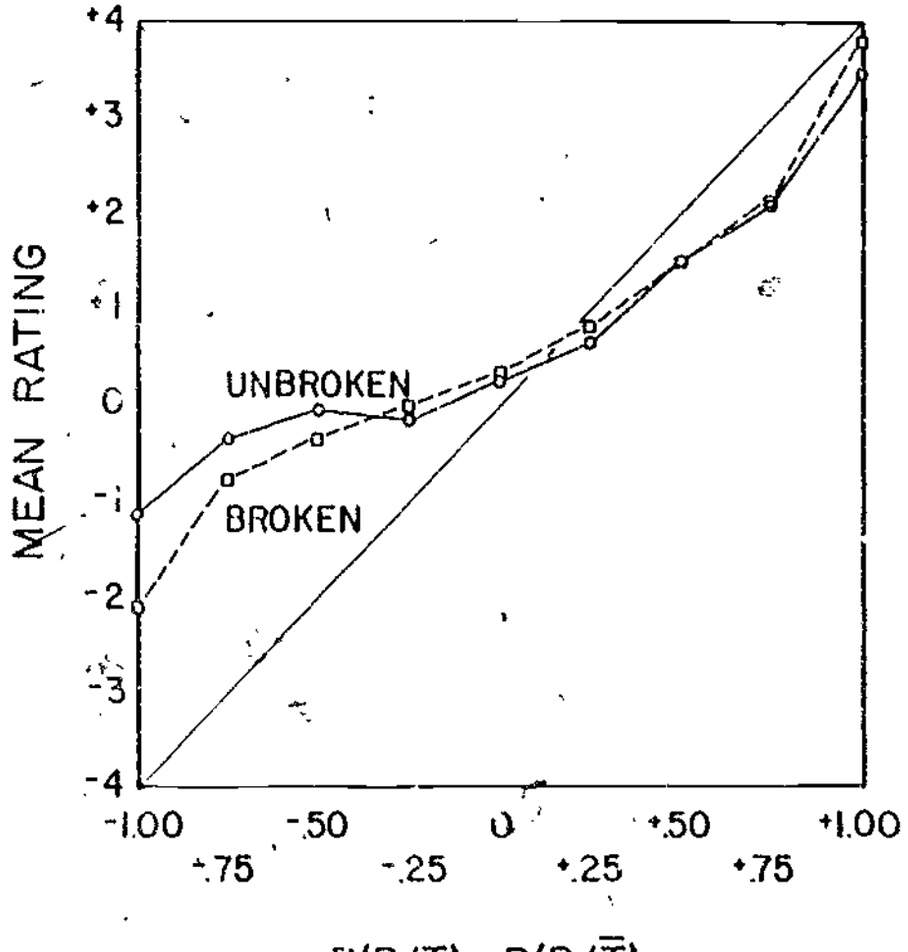

+1.00
+.75
+.50
+.25
0 = P(B/T)− P(B/T̄)
1.00
.75
.50
.25
0
P(B/T)
-.25
-.50
-.75
-1.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
P(B/T̄) → 0 .25 .50 .75 1.00
1.00
.875
.750
.625
.500 = P(B)
.125
.250
.375

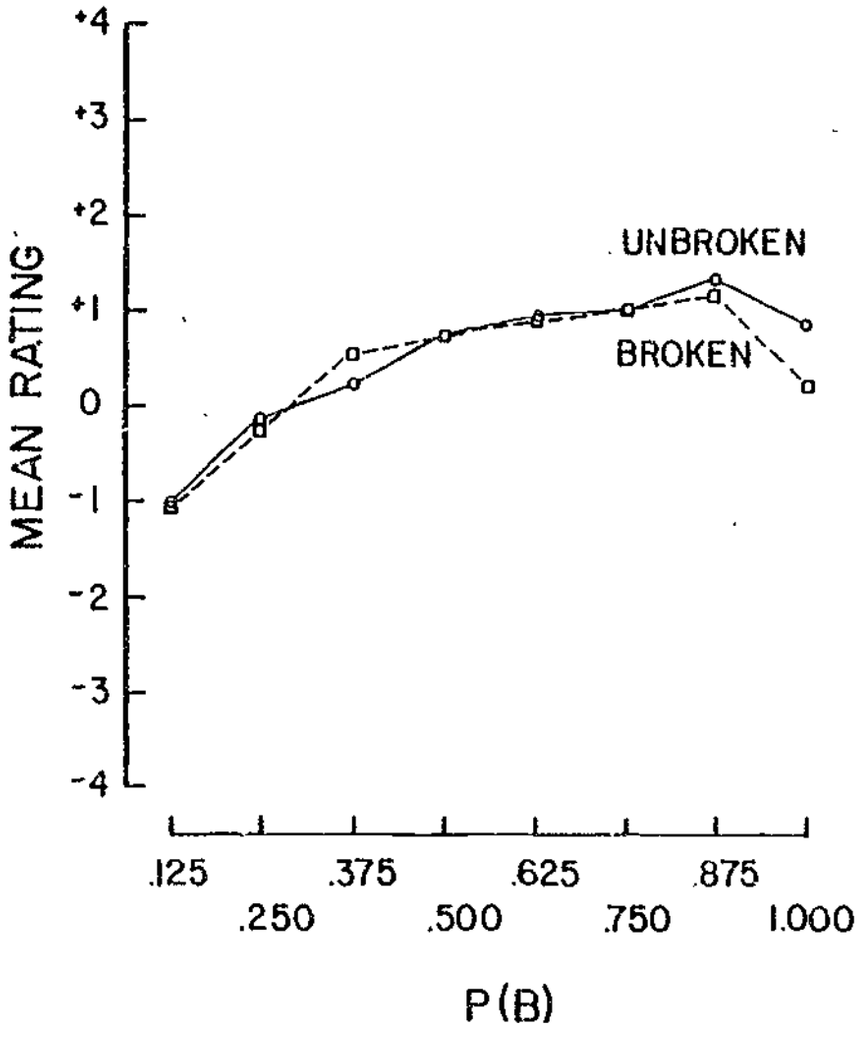

MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
WITHIN
BETWEEN
-1.00
-.75
-.50
-.25
0
+.25
+.50
+.75
+1.00
P(B/T)−P(B/T̄)

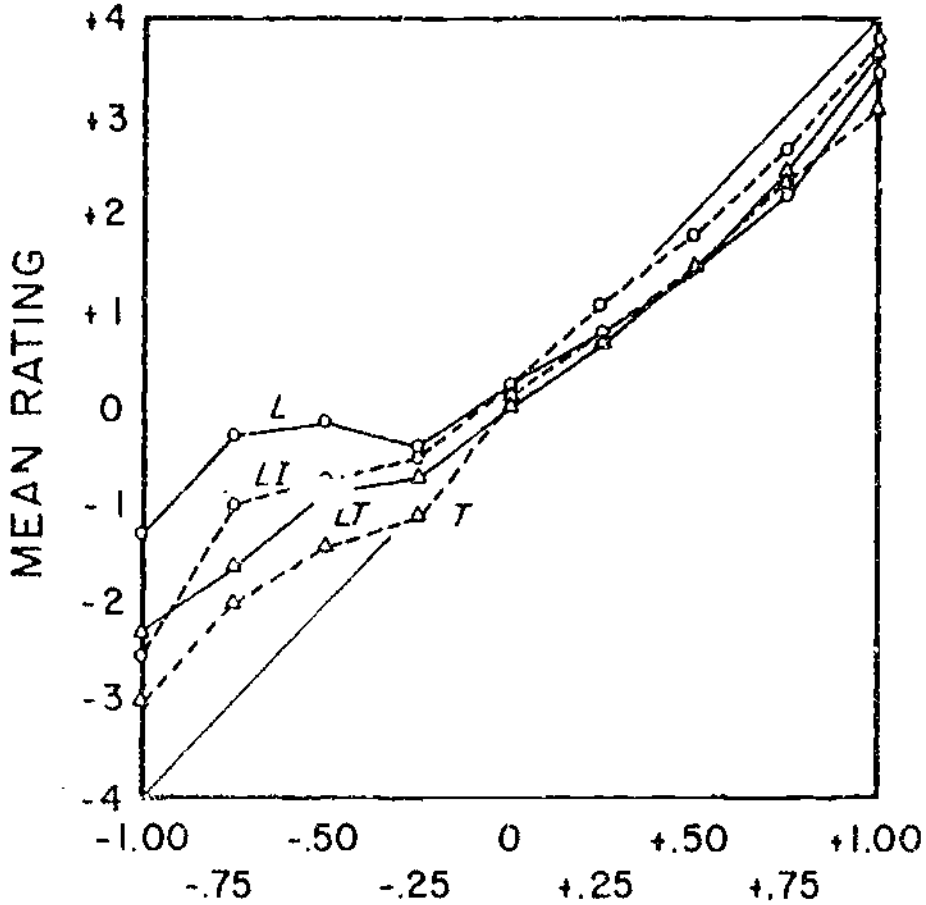

MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
WITHIN
BETWEEN
.125
.250
.375
.500
.625
.750
.875
1.000
P(B)

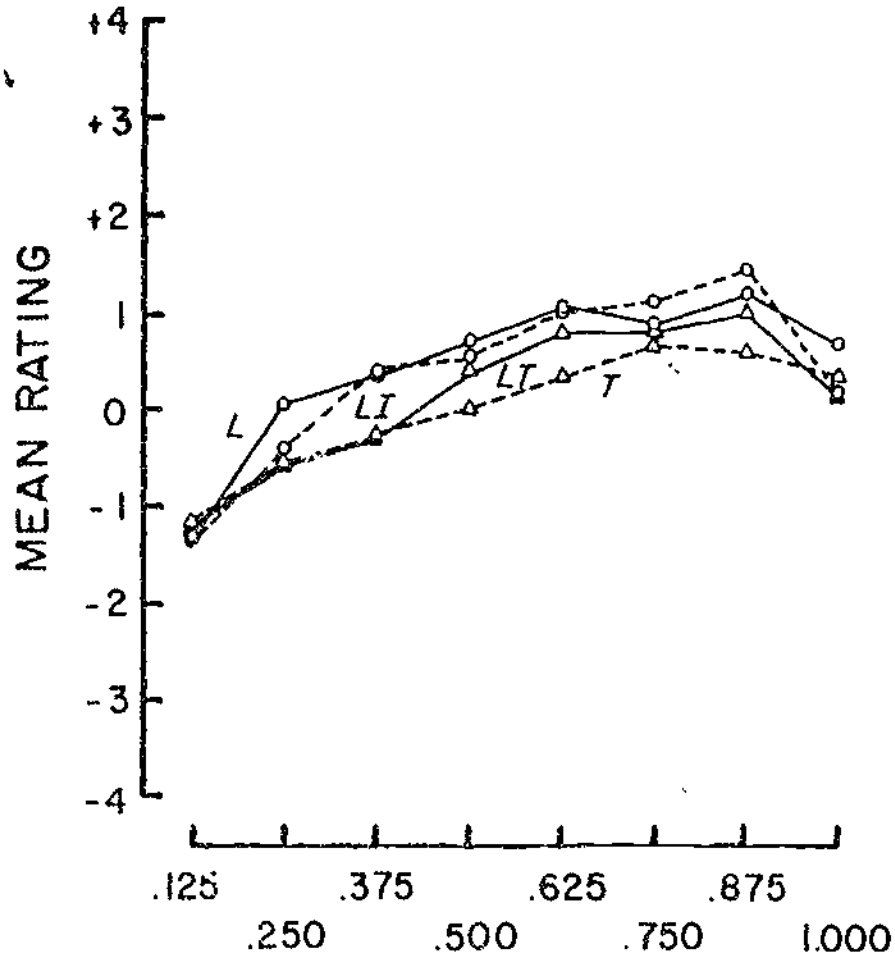

MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
LINE
TABLE
-1.00
-.75
-.50
-.25
0
+.25
+.50
+.75
+1.00
P(B/T)-P(B/T̄)


MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
LINE
TABLE
.125
.250
.375
.500
.625
.750
.875
1.000
P(B)


MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
UNBROKEN
BROKEN
-1.00
-.75
-.50
-.25
0
+.25
+.50
+.75
+1.00
P(B/T) − P(B/T̄)


MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
UNBROKEN
BROKEN
.125
.250
.375
.500
.625
.750
.875
1.000
P(B)


MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
L
LI
LT
T
-1.00
-.75
-.50
-.25
0
+.25
+.50
+.75
+1.00
P(B/T) - P(B/T̄)


MEAN RATING
+4
+3
+2
+1
0
-1
-2
-3
-4
L
LI
LT
T
.125
.250
.375
.500
.625
.750
.875
1.000
P(B)