Memory & Cognition
1988, 16 (3), 271-280

Prediction and judgment as indicators of sensitivity to covariation of continuous variables

ARNOLD D. WELL, SUSAN J. BOYCE, ROBIN K. MORRIS,
MAKIKO SHINJO, and JAMES I. CHUMBLEY
University of Massachusetts, Amherst, Massachusetts

The present study used both judgments of strength of relationship and measures of the ability to predict one variable from another to assess subjects' sensitivity to the covariation of two continuous variables. In addition, one group of subjects judged strength of relationship after merely observing the presentation of 60 pairs of two-digit numbers, and a second group made strength judgments after being actively engaged in predicting one member of a pair when given the other. The prediction and judgment data provide different pictures of subjects' sensitivity to covariation. The subjects were quite poor at estimating strength of relationship but, by some measures, good at predicting one variable from another. Judgments were not strongly influenced by whether subjects had previously engaged in overt prediction. The implications of these results for the literature on covariation estimation are discussed.

The ability to detect relationships between events in the environment and to use knowledge about these relationships to make predictions has come to be regarded as an important component of human intelligence. As Crocker (1981, p. 272) pointed out, "knowing whether events are related, and how strongly they are related, enables individuals to explain the past, control the present, and predict the future."

The large literature that is concerned with people's ability to judge the degree to which imperfectly related events covary has often been characterized as indicating that people are poor at assessing covariation (e.g., Nisbett & Ross, 1980). Most research in the field has dealt with binary variables (e.g., presence or absence of symptom and presence or absence of disease), so that all cases fall into one of the cells of a 2×2 table. Empirical studies and theoretical analyses have been aimed at exploring reasons why performance is often poor and at delineating circumstances under which subjects tend to be more or less accurate according to some statistical measure of relationship (for reviews, see Alloy & Tabachnik, 1984; Crocker, 1981).

The ability to estimate covariation between variables that can take on more than two values has also been characterized as being quite poor (e.g., Nisbett & Ross, 1980), despite the fact that few estimation studies have employed nonbinary variables and the fact that the older studies cited as providing evidence of conservatism (Beach & Scopp, 1966; Erlick & Mills, 1967) are difficult to compare with more recent ones.[1] The dependent variable used by Beach and Scopp (1966) was a rating of degree of confidence that the relationship was positive or negative, whereas subjects in the Erlick and Mills (1967) study made more than 1,000 covariation judgments over 10 sessions.

Jennings, Amabile, and Ross (1982) had subjects make covariation judgments for several types of "theory-free" bivariate distributions in which the product-moment correlation (r) varied between 0 and 1. In one condition, after briefly studying 10 number pairs, subjects first judged whether the relationship between members of pairs was positive or negative, then indicated their subjective impression of the strength of the relationship. Judgments of strength were made by placing an "X" on a 100-point rating scale on which the end points were labeled *no relationship* and *perfect relationship*. Estimates were extremely variable, and moderately large correlations were barely detected. The average ratings tended to be characterized by the function $100(1 - \sqrt{1-r^2})$, so that objective correlations as large as .4 were rated in the lower 10 to 15 points of the 100-point scale and the average ratings did not reach the midpoint of the scale until the objective correlation approached a value of .85.

However, as Crocker (1981) pointed out, it is not clear exactly what conservatism means in a covariation judgment task. It is not obvious that r should serve as the normative criterion, or that r is in any sense a more meaningful index of strength of linear relationship than r^2 or even $1 - \sqrt{1-r^2}$. A similar point was made by Wright and Murphy (1984), who found that subjects' estimates (again on a 100-point scale) were more linearly related to a measure of correlation suggested by Cleveland (1979) that is less sensitive to outliers than is the standard Pearson r. They also found that relationships tended to be perceived as stronger when cover stories encouraged the expectancy of a strong, as opposed to a weak, relationship, but that

The authors would like to thank Axel Buchner, R. Kevin Stone, and Laura Ringey for their assistance in data collection. Requests for reprints should be sent to either Arnold Well or James Chumbley, Department of Psychology, University of Massachusetts, Amherst, MA 01003.

even cover stories that encouraged the expectancy of a weak relationship had beneficial effects that were attributed to making the problems meaningful. An interesting procedural difference between the Jennings et al. (1982) and the Wright and Murphy (1984) studies is that although subjects in both were instructed to estimate strength of relationship using a rating scale, the instructions in the latter explicitly explained the concept of the strength of a relationship as "how well one could predict the score on one variable from the score on the other" (p. 306).

Another way to investigate subjects' sensitivity to the nature of the relationship between two variables is to forgo the use of a rating scale and instead observe how information about one variable is used in making predictions about the second. Admittedly, prediction performance depends both on the subject's ability to register information about the nature of the relationship between the two variables and on the ability to use this information in making predictions (Beach & Scopp, 1966). Nonetheless, if subjects are able to achieve a high level of prediction performance when given two variables that are linearly but imperfectly related, it would be difficult to argue that they are insensitive to the covariation between the variables, even if they are unable to provide explicit judgments that closely reflect some normative criterion such as the correlation coefficient. Given the difficulties associated with using rating scales and deciding what to use as the normative criterion against which to compare ratings of covariation, prediction measures may provide useful information about the ability to assess covariation.

A large body of research using a linear regression model to conceptualize prediction and judgment tasks has accumulated during the past 2 decades (see, e.g., Brehmer, 1973; Naylor & Domine, 1981; Sniezek, 1986). Typically, in the so-called *cue probability learning task*, on each of a number of trials, the subject is provided with the value(s) of one or more cues or predictor variables (X) and is required to generate a prediction (P) of an uncertain future event or criterion (Y). After the prediction has been generated, the subject is usually provided with feedback concerning the correct criterion value. During the course of the experiment, the subject acquires an understanding of the relationship between the predictor variable(s) and the criterion, which is reflected in improved predictions.

In a subset of this literature, the task is to predict the criterion from a single predictor variable and is referred to as *single-cue probability learning* (SPL). Data provided by SPL studies in which the predictor and criterion have an imperfect linear relationship would seem to be relevant to the issues of covariation assessment. Although the task is usually abstract or content-free (i.e., no cover story is presented and the variables are not labeled), performance is often quite good according to some of the measures used to evaluate prediction performance. However, with only a few exceptions (Lane, Anderson, & Kellam, 1985; Malmi, 1986), this literature seems to have been virtually ignored by researchers concerned with covariation assessment.

If prediction performance is to be used to assess sensitivity to covariation, what measures should be used to evaluate the predictions? According to the usual least squares criterion, optimal linear predictions of Y based on X should all lie exactly on a regression line with slope $b_{YX} = r_{XY}(S_Y/S_X)$, where S_X and S_Y are the standard deviations of X and Y and r_{XY} is the correlation. Therefore, one index of a subject's performance in a linear prediction task is the slope of the line obtained by regressing the subject's predictions on the values of X; that is, $b_{PX} = r_{PX}(S_P/S_X)$. To the extent that b_{PX} exceeds b_{YX}, the subject may be characterized as extreme in his/her predictions because a given change in the value of the predictor tends to result in larger changes in the predictions than in the criterion values. Conversely, to the extent that b_{PX} is less than b_{YX}, the subject may be characterized as conservative (see, e.g., Brehmer, 1976).

Other measures used in the SPL literature provide an index of the extent to which the subject consistently uses a linear prediction strategy. One such measure is r_{PX}, the correlation between the values of the predictor variable and the subject's predictions. A second consistency measure is S^2_{PX}/S^2_{YX}, where S^2_{YX} is the variance of estimate for the regression of Y on X (i.e., a measure of the variability about the regression line) and S^2_{PX} is the corresponding measure for the regression of the subject's predictions on X. If a subject generated predictions that had a perfect linear relationship with X, the value of the ratio would be 0; if the distribution of predictions simply matched the distribution of the criterion, the ratio would take on a value of 1.

An additional measure used to assess prediction is r_{PY}, the correlation between the subject's predictions and the values of the criterion. The maximum possible value this measure (termed the *achievement* measure in the SPL literature) could take on if the subject used the information about X to generate optimal linear predictions of Y would be r_{XY}.

Malmi (1986) used a prediction task to study what he called "intuitive covariation estimation." In different experiments, stimuli were pairs of numbers, pairs of lines of variable length, and word-line pairs. The use of arithmetic strategies was discouraged by using nonnumerical stimuli and by using rapid presentation when numerical stimulus pairs were displayed. In Malmi's experimental procedure, subjects were first presented with a large number of X-Y pairings. Subjects were then shown a small number of test stimuli that had not appeared earlier and were asked to predict what the other member of the pair should be for each test stimulus, by generating either a number or a line length. Measures of covariation estimation were based on prediction performance. Malmi's conclusions were based on two dependent variables: (1) the b_{PX} measure mentioned earlier and (2) a measure he termed the "subjective correlation coefficient" that was inferred from the prediction data. Malmi's description of

this latter measure is unclear, but we believe the measure to be approximately $r_{PX}(S_P/S_Y)$. Performance was characterized as good, except when the stimuli were numerical and the sign of the correlation between the stimulus variables was negative.

We agree with Malmi's (1986) attempt to gain additional information about sensitivity to the relationship between two variables by considering information about prediction performance, although we are not sure it is appropriate to refer to Malmi's prediction-based performance measures as indices of covariation "estimation." However, we believe that it is premature to conclude that subjects perform well if forced to assess covariation "intuitively" but perform poorly if allowed the opportunity to engage in strategies of one sort or another. The SPL literature contains studies in which subjects' use of strategies for predicting a numerical criterion was not discouraged (and was even sometimes encouraged by instructions) and in which performance would be considered to be good according to prediction measures (e.g., Brehmer, 1973, 1974).

Given that Malmi (1986) did not ask subjects to estimate strength of relationship, but rather employed measures derived from predictions, it is difficult to say whether the subjects in his study (in which performance was characterized as good) had any better understanding of the relationship between the two variables than those in, say, the Jennings et al. (1982) study (in which performance was characterized as poor). Before any definitive conclusions about strategy can be made, it is necessary to obtain both estimation and prediction data from the same subjects, so that the different measures of performance can be compared. Also, it is possible that the very act of being involved in a prediction task improves understanding of the relationship between the predictor and criterion, so that subjects who predict one variable from another may also be able to provide more accurate judgments of strength of relationship than those who simply observe the paired values.

In the present study, one group of subjects participated in a prediction task, following which they estimated the strength of the relationship between the two variables. A second group of subjects observed the same stimulus pairings without predicting, and then estimated the strength of relationship.

METHOD

Subjects

Forty undergraduate volunteers at the University of Massachusetts at Amherst received extra credit in psychology courses for participation.

Materials

Each subject was presented with three sets of paired two-digit numbers that had correlations of .9, .6, and .1. The single cover story that was used to provide context incorporated variables that were familiar to subjects (commuting distance and work efficiency for employees of three different companies) but did not evoke any strong prior belief about strength of relationship. The cover story is presented in Appendix A.

Each stimulus set consisted of 60 X-Y (distance-efficiency) pairings. For each set, the standard deviation was 10 for both the X and Y scores, and the means were 20 and 75 for X and Y, respectively. The sets of number pairs were obtained by first sampling sets of uncorrelated scores, z_X and z_e, from the unit normal distribution. A set of scores z_Y, having the desired correlation, r, with z_X was obtained using $z_Y = rz_X + \sqrt{1-r^2}\, z_e$, and then z_X and z_Y were transformed to have the appropriate means and standard deviations. The values of z_X and z_Y were constrained to lie between ± 2.0 and ± 2.3, respectively, keeping commuting distance greater than 0 and the work efficiency measure (which was expressed in terms of percentage of projects completed) less than 100. The order in which the number pairs were presented was randomized separately for each subject and each stimulus set. The correlation was kept within .002 of the desired value for each set of 60 pairs as well as separately for the first 30 and last 30 pairs in each set.

Design

The subjects were randomly assigned to either the prediction group (which engaged in prediction and then judged strength of relationship) or the observation group (which merely observed the number pairs before judging). The other between-subjects variable was correlation order. Half of the subjects received the order .9, .6, .1, and the other half, .1, .6, .9.

Procedure

All subjects participated individually at a computer terminal controlled by a North Star Horizon microcomputer. An experimental session lasted approximately 30 to 45 min.

The subjects began the session by reading instructions that included a very brief explanation of the concept of relationship (the instructions are presented in Appendix B). Instructions on the screen then indicated that stimulus presentation would be self-paced and that pressing a key would result in presentation of an X-Y pair on the screen. For the observation group, both members of each pair were presented simultaneously. Each keypress resulted in another pair being added to the display until 30 pairs had been presented. After the subjects had studied the display for as long as they wished, the display was cleared and the next 30 pairs were presented in the same fashion. At this point, the subjects were allowed to view a display containing all 60 pairs until they felt ready to evaluate the strength of relationship. The subjects were then presented with a display consisting of a rating scale labeled 0 (*no relationship*) on the left and 100 (*perfect relationship*) on the right and asked to enter a number from 0 to 100 that most accurately represented their judgment of the strength of relationship between X and Y. After providing their judgment, the subjects moved on to the second and third stimulus sets (i.e., distance-efficiency data from the second and third companies).

For the prediction group, presentation of the first 30 number pairs took place in exactly the same fashion, except that each Y value was displayed 2 sec after the corresponding X in order to encourage subjects to generate predictions of Y. For the second 30 pairs, the subjects were required to make predictions, using information about the pairs they had already seen. A value of X was displayed, the subject typed in his or her prediction (P, which also remained on the screen), and then the value of Y was displayed between the X and P values. Following the 30 prediction trials, the screen was cleared and replaced by a display of all 60 pairs (without predictions). Judgments of the strength of relationship were then obtained exactly the same way as in the observation group.

RESULTS

Judgments of Strength of Relationship

The judgment and prediction results are summarized in Table 1. In general, judgments of strength of relation-

Table 1
Summary Judgment and Prediction Data

	Objective Correlation (r_{XY})					
	.1		.6		.9	
	M	*SEM*	*M*	*SEM*	*M*	*SEM*
Strength Judgments						
Prediction Group	29.05	5.37	41.65	3.94	61.35	3.69
Observation Group	30.35	5.16	32.95	5.13	57.05	5.95
Prediction Measures						
b_{YX}	.1		.6		.9	
a_{YX}	72		57		48	
b_{PX}	.07	.06	.65	.04	.81	.03
a_{PX}	74.56	2.03	55.83	1.50	50.60	1.09
S_Y	10		10		10	
$S_{Y'}$	1		6		9	
S_P	7.28	.47	8.79	.40	9.15	.35
S^2_{PX}/S^2_{YX}	.48	.05	.51	.07	.91	.11
r_{PX}	.08	.09	.74	.03	.89	.05
r_{PY}	−.04	.04	.38	.04	.78	.05

Note—$N = 20$. *The mean value of the predictor variable (X) was 20 and the standard deviation was 10.

ship increased with degree of covariation between X and Y, but not in a fashion that closely reflected r, r^2, or $1 - \sqrt{1-r^2}$. When the judgment data were subjected to an analysis of variance (ANOVA) with two between-subjects factors, condition (prediction, observation) and correlation order (high, medium, low; low, medium, high), and the within-subjects factor correlation (.9, .6, .1), only the correlation main effect was significant [$F(2,72) = 28.63$, $MSe = 327.73$, $p < .00001$].

Although neither the condition main effect nor any of the interactions involving the condition factor approached significance, there was some suggestion from the judgments that prediction group subjects discriminated better between medium and low correlation pairs than did observation group subjects. For the observation group, judgments of medium correlation pairs were significantly smaller than those of high correlation pairs [$t(19) = 4.59$, $p < .001$], but did not differ from judgments of low correlation pairs ($t < 1$). For the prediction group, judgments of medium correlation pairs were significantly different from judgments of both high and low correlation pairs [$t(19) = 4.33$, $p < .001$, and $t(19) = 2.61$, $p < .02$, respectively].

There was a considerable amount of variability in how the subjects used the 100-point judgment scale. The judgment range (i.e., the difference between a subject's largest and smallest judgments) tended to be quite small, averaging 37.8 and 38.2 for the observation and prediction groups, respectively, and this range varied appreciably from subject to subject (standard deviations of 25.9 and 22.1). Moreover, the subjects who used only a small part of the scale tended to vary considerably in what part of the scale they chose to use. The means of the three judgments made by each subject had standard deviations of 18.9 and 15.0 for the observation and prediction groups, respectively.

Because of the variability in the use of the judgment scale across subjects, we were concerned that the means and standard deviations of the subjects' judgments may not have provided an accurate indication of their ability to discriminate differences in covariation. Therefore, we also analyzed ordinal properties of the subjects' judgments, noting the presence of judgment *reversals*. A reversal was counted if the estimated strength of relationship was the same or larger for a lower correlation stimulus set than for a higher correlation set. Although there were no significant group differences, prediction group subjects tended to exhibit fewer judgment reversals. Fourteen prediction and 11 observation subjects had no judgment reversals. The remaining 15 subjects had a total of 24 reversals: 11 low-medium reversals (8 in the observation group), 7 medium-high reversals (4 in the observation group) and 6 low-high reversals (4 in the observation group).

Prediction Measures

The indices generally used to evaluate prediction performance indicated that the subjects were quite sensitive to the relationship between X and Y. The first measure discussed is b_{PX}, the slope of the regression line fitting the subjects' predictions. If b_{PX} is nonzero, it could be argued that the subjects used information about X in predicting Y. If b_{PX} is similar in value to b_{YX}, the slope of the optimal regression line, it could be argued that, on the average, the subjects used information about X appropriately. The obtained slope, b_{PX}, was very similar to b_{YX} (see Table 1) and was quite stable across subjects, with 95% confidence intervals of $.07 \pm .13$, $.65 \pm .08$, and $.81 \pm .06$, as opposed to the optimal values of .1, .6, and .9. There was a total of three reversals, all between high and medium correlation stimuli. The ANOVA for b_{PX} showed a large main effect due to correlation [$F(2,36) = 71.48$, $MSe = .043$, $p < .00001$]. There was also a main effect of correlation order [$F(1,18) = 9.22$, $MSe = .045$, $p < .01$], suggesting that subjects tended to use information about X more extensively in predicting Y when this information had been more useful in the previous conditions. The mean values of b_{PX} for the high, medium, and low correlation conditions were .86, .71, and .21, respectively, when the high correlation came first. The corresponding values when the low correlation condition came first were .76, .60, and −.07.

Not only was b_{PX} extremely similar to b_{YX}, the optimal slope, but the mean values of the subjects' predictions ($\bar{P}$) were very close to the means of Y distributions ($\bar{Y}$) for all three stimulus sets (74.90, 75.47, and 76.76 for the high, medium, and low correlation conditions, respectively, as opposed to the mean Y value of 75). Therefore, the intercepts of the regression lines that characterized the subjects' predictions, $a_{PX} = \bar{P} - b_{PX}\bar{X}$, were almost identical to the intercepts (a_{YX}) of the optimal regression lines (74.56, 55.83, and 50.60 for the low, medium, and high objective correlation conditions, as opposed to the optimal values of 72, 57, and 48).

On the other hand, the subjects' predictions were more variable than would have been the case had they been generated using the optimal linear regression equation. Optimal least squares predictions (Y') fall exactly on the regression line, and therefore regress toward the mean ($\bar{Y}$) as r_{XY} decreases, yielding standard deviations of $S_{Y'} = r_{XY}S_Y$. Although the standard deviations of subjects' predictions (S_P) did vary as a function of correlation [$F(2,36) = 12.49$, $MSe = 1.47$, $p < .0002$], they did not decline as much as $S_{Y'}$ when the value of the correlation became smaller (see Table 1). The consistency ratio used in the prediction literature, S^2_{PX}/S^2_{YX}, took on values of .48 and .51 for the low and medium correlation conditions, indicating that although there certainly was not zero variability about the subjects' prediction regression lines, there was a good deal more consistency in subjects' predictions than there was in the values of the criterion variable.

Two other prediction measures that we analyzed were r_{PX}, a measure of the extent to which subjects' predictions are linearly related to X, and r_{PY}, the correlation between subjects' predictions and the criterion value. Totally consistent adherence to a linear prediction strategy would result in r_{PX}'s having a value of 1, even if r_{XY} was very small. The values of r_{PX} were quite large (.74 and .89) in the medium and high correlation conditions and very small (.08) in the low correlation condition. This indicates that subjects' predictions were quite consistent with a linear strategy in those conditions in which there was a linear predictability and were not consistent in the condition in which there was virtually no linear predictability.

The r_{PY} measure suggests that the subjects were good at taking advantage of the degree of predictability that existed in the stimulus pairs and made better predictions as the degree of predictability between X and Y increased. There were no reversals on r_{PY} for any of the 20 subjects in the prediction group: in every case, r_{PY} was larger for the high correlation stimulus set than for the medium correlation set, and larger for the medium correlation set than for the low correlation set. For both r_{PX} and r_{PY}, ANOVAs showed a large main effect of correlation [both with values of $F(2,36) > 76.6$, $p < .00001$], with no other effect approaching significance.

DISCUSSION

One major purpose of the present study was to use both judgment and prediction data to investigate the extent to which subjects are sensitive to the covariation between two variables. The two types of data seem to provide conflicting pictures of this sensitivity. The subjects were quite poor at generating global judgments of strength of relationship, but good at predicting one variable from another, even though they started making predictions after seeing only 30 data pairs and judged strength of relationship after all 60 pairs had been presented. The second major purpose of the study was to investigate whether engaging in the task of predicting Y from X, as opposed to merely observing X-Y pairs, had an effect on subsequent judgments of strength of relationship. Judgments were not strongly influenced by whether or not the subjects had engaged in overt prediction. There was, however, a suggestion of a small advantage in discriminating low and medium correlation pairs for the prediction group relative to the observation group. If the effect is real, it could be due to the successive presentation format as well as to the act of generating predictions.

Our judgment data were roughly consistent with those obtained by Jennings et al. (1982). Judgments were extremely variable, and, although we used only three correlation values, there seemed to be less discrimination at the low end of the correlation scale than at the high end. The mean judgments were considerably larger for $r_{XY} = .1$. and .6 in the present study than in that of Jennings et al., but this is due to the fact that our subjects judged only strength of relationship, whereas Jennings et al.'s subjects first judged whether the relationship was positive or negative and then how strong it was. The large number of negative judgments (reading from Figure 1 of the Jennings et al. paper, one-quarter of the judgments were more negative than -38 for $r_{XY} = +.1$ and more negative than -21 for $r_{XY} = .6$) reduced the averages at the low end of the correlation scale.

As indicated earlier, it is not clear what normative measure should be used to evaluate judgments of strength of relationship. However, the subjects' ability to provide numerical judgments of covariation cannot be considered to be very good, no matter what reasonable criterion might be chosen, given the large amount of between-subject variability and the fact that the judgments for 15 of the 40 subjects contained at least one reversal.

According to several standard measures of prediction performance, the subjects seemed to be quite sensitive to degree of covariation. In particular, the linear regression equations that characterize the subjects' predictions were extremely similar to the optimal regression equations for all three objective correlations. Moreover, the so-called "achievement" measure, r_{PY}, although smaller than r_{XY}, varied systematically with it and showed no reversals for any of the 20 subjects. Performance was not optimal, however, in that there was more variability in the subjects' predictions than would have been the case had the optimal linear prediction strategy consistently been used. Possible reasons for this variability are discussed later.

Given that people seem more sensitive to covariation according to prediction measures than according to global numerical judgments, it is important that comparisons among studies be based on the same indices of performance. One cannot conclude very much about the effects of a particular type of stimulus presentation when differences in presentation are completely confounded with differences in response measures. Malmi (1986) used prediction measures to conclude that subjects were able to assess covariation adequately. He attributed the good performance to the fact that his subjects were engaged in an "intuitive" mode of functioning, inasmuch as his stimulus presentation procedures discouraged the use of strategies that may have occurred in studies such as that

of Jennings et al. (1982) in which poor performance was reported. However, it is not appropriate to draw conclusions about the relative efficacy of Malmi's and Jennings et al.'s procedures for presenting stimuli, because they used different measures to assess performance. Our materials were not the same as those used by either Malmi or Jennings et al. Nonetheless, using the same subjects and stimulus presentation, we obtained prediction performance that was at least as good as Malmi's and estimation performance that was arguably as bad as that obtained by Jennings et al.

Why Might Prediction Measures Indicate More Sensitivity to Covariation than Global Judgments of Strength of Relationship?

There are a number of possible reasons why the sensitivity to the relationship between variables that is indicated by the prediction data is not reflected in subjects' judgments of strength of relationship. Prediction may be a better defined, less ambiguous task than that of producing global judgments of strength. Subjects have no difficulty understanding instructions to try to predict as accurately as possible and seem comfortable with the notion of making predictions.

On the other hand, subjects are not accustomed to making explicit judgments of strength of relationship, and, in the current study, the subjects frequently seemed uncomfortable with the idea of using a number from 0 to 100 to represent the degree of strength. They also varied considerably with respect to how much of and what part of the 100-point scale they used. Moreover, there are several quite different interpretations of what may be meant by the term *strength of relationship*, and judgments of strength have been elicited in different ways. Therefore, it is not clear what features of the relationship subjects attend to when making these judgments and whether all subjects attend to the same features.

Two reasonable but quite different interpretations of the strength of relationship between X and Y are (1) how much, on the average, one variable changes as the other changes and (2) how predictable one variable is from the other. These different interpretations lead to consideration of different aspects of the relationship between X and Y. If X is the independent variable and Y the dependent variable, the slope of the regression line of Y on X, $b_{YX} = r_{XY}(S_Y/S_X)$, is a measure of how much, on the average, Y changes when X changes. In contrast, measures of how predictable Y is from X are based on how well the regression line fits the data. When Y is predicted from X, the variability about the regression line is measured by $S_{YX} = S_Y\sqrt{1-r^2}$, the so-called *standard error of estimate*, or its square, the *variance of estimate*. Because the amount of variability about the regression line indicates the degree of nonpredictability, complementary measures, such as $S_Y(1 - \sqrt{1-r^2})$ or $S_Y^2 r^2$, would seem to be reasonable indices of predictability.

When some statistics books that are primarily concerned with regression or causal modeling (e.g., Achen, 1982; Hanushek & Jackson, 1977) refer to strength of relationship between an independent variable X and a dependent variable Y, they mean the regression slope coefficient, b_{YX}; if there is a causal relationship, it is the slope that is regarded as the measure of *causal power*. In contrast, these authors assert that r and r^2 are not good indices of strength because they are *composite* measures and because they are *sample-specific*.

The correlation coefficient can be thought of as a composite measure[2] that incorporates different aspects of the linear relationship between X and Y—namely, the variances of X and Y, S_X^2 and S_Y^2, and the slope and variability about the optimal regression line, b_{YX} and S_{YX}^2—as follows:

$$r_{XY}^2 = \frac{b_{YX}^2 S_X^2}{S_Y^2} = \frac{b_{YX}^2 S_X^2}{b_{YX}^2 S_X^2 + S_{YX}^2}.$$

It is possible that it is more adaptive to be sensitive to separate components of a relationship, such as rate of change and predictability, than it is to be directly sensitive to a composite measure, such as the correlation coefficient. If so, perhaps some of the high degree of variability that seems to characterize subjects' judgments of strength occurs because different subjects attend to different aspects of the relationship or combine them in different ways when instructed to make "strength" judgments.

Also, the correlation coefficient mixes the information about the different components of a linear relationship in a way that makes its value very dependent on the variances of the variables. The correlation coefficient is termed a sample-specific measure because, for a linear relationship with given values of b_{YX} and S_{YX}, the value of r obtained from a sample will depend critically on the range of X values that are sampled in a way that is not true of either b_{YX} or S_{YX}. This makes it difficult to compare relationships across samples if one uses correlations instead of regression parameters.

In referring to r and r^2, Achen (1982) commented,

> The fact that a Pearson r (or a gamma, phi, standardized beta, or any other correlational measure) depends in an important way on the variance of the variables involved makes comparisons meaningless in general. Different correlational measures depend on the variance in different ways, but the solution is not to find the one that captures the medieval essence of correlation, but rather to abandon them all.... It neither measures causal power [i.e., the slope] nor is it comparable across samples.... It makes little sense to base decisions on a statistic that for most social science applications measures nothing of serious importance. (p. 61)

The advice that Achen (1982) directed to scientists attempting to understand and model social processes may also be applicable to organisms attempting to understand and model their environments. The correlation coefficient would not be a very useful measure for understanding cur-

vilinear relationships and may be less useful than the slope and measure of fit for understanding linear ones.

In fact, at the present time we know little about how the different components of the relationship between X and Y influence subjects' judgments of strength. It is possible that subjects are not only influenced by the slope and fit in a way other than the model suggested by the correlation coefficient, but that what they attend to depends on the details of the instructions and even on the nature of the cover story.

Different investigators may have elicited judgments of strength using instructions that do not seem to call attention to exactly the same features of the relationship. Jennings et al. (1982) asked subjects to judge how strong the relationship was; Wright and Murphy (1984) asked for a rating of the strength of relationship, but defined strength as "how well one could predict the score on one variable from the score on the other" (p. 306); and Lane et al. (1985) asked for ratings on a scale of 0 to 100, where "zero means no relationship and 100 means a perfect linear relationship" (p. 642). These instructions may mean different things to different subjects and, taken literally, may not ask for a judgment of the correlation coefficient.

The cover story may also help determine which aspects of the relationship are attended. If, for example, X is clearly an independent variable and Y is a dependent variable, it seems reasonable that perceived strength may depend more directly on the rate of change of Y with X, b_{YX}, than on r_{XY}. On the other hand, if X and Y are symmetric in the sense that neither is more likely to be considered an independent or dependent variable than the other, perceived strength may depend both on b_{YX} and on b_{XY}, the rate of change of X with Y, and may therefore more closely reflect r_{XY}, which can be thought of as a kind of average of the rate of change of one variable with the other (it is the geometric mean of b_{YX} and b_{XY}).

We know of only a single attempt to investigate the effects of different components of the relationship between two variables on judgments of strength of relationship. Lane et al. (1985) investigated the effects of the variance of X, the regression slope, and the "error variance" (i.e., the variability about the regression line, $S^2_{Y\cdot X}$) on judgments of relatedness. They observed that different combinations of these components could lead to different judgments of relatedness, even when r_{XY} was kept constant, and concluded that people are influenced more by error variance than by either slope or variance of X relative to how these factors contribute to the size of the correlation coefficient. Unfortunately, this very interesting conclusion was based almost entirely on judgments about scatter diagrams. When the same data were presented in tables, judgments of relatedness were much smaller (possibly because instructions to judge the extent to which there was a "perfect linear relationship" were easier for naive subjects to apply to the graphic than to the tabular format). There was some tendency for the same kinds of effects to occur for tabular presentation, but the effects were much compressed and were not significant. Clearly, more work of this type should be done.

Also, if in some situations, strength judgments are based on perceived predictability, it is important to determine how subjects evaluate different-sized errors in prediction. Are subjects sensitive to the absolute size of their prediction errors or to their relative size? And, if sensitive to relative size, relative to what? We have pilot data that suggest that subjects are partly sensitive to the absolute size of prediction errors rather than only to their z scores. We used different cover stories for which the distributions of X and Y scores varied in both mean and variance. Although we have not collected data in enough conditions to make definitive statements about how global strength judgments are influenced by systematic manipulations of slope, variance of X and Y, and error variance, when subjects made a series of predictions and then judged strength of relationship, the judgments for a given objective value of r_{XY} tended to be higher for small values of S_{YX} and S_Y. If subjects completely compensated for variability (effectively perceiving quantities in terms of their z scores), only the objective value of the correlation coefficient, and not how it was varied, should matter.

Some Observations from the Single-Cue Probability Learning Literature

As mentioned in the introduction, some of the findings in the SPL literature (e.g., Brehmer, 1973; Kuylenstierna & Brehmer, 1981; Naylor & Domine, 1981) are relevant to assessment of covariation. In the usual SPL task, the subject is presented with the predictor variable, makes a prediction and then receives as feedback the true value of the criterion. Stimuli are constructed so that the relationship between the predictor and the criterion is "statistical"; that is, the criterion is not completely predictable from the predictor. Almost always, the task is completely abstract; that is, stimulus pairs are presented without cover story or variable labels. Sometimes both the predictor and the criterion are numbers (e.g., Naylor & Domine, 1981; Slovic, 1974), but often the predictor is a nonnumerical quantity such as the length of a line (e.g., Brehmer, 1973). The emphasis is on how prediction performance improves as the subject infers and then learns how to use the functional rule that characterizes the relationship. There is concern with how the course of learning proceeds when there are different kinds of functional relationships (e.g., positive linear, negative linear, U-shaped, J-shaped) and different degrees of predictability. In many studies, it has been found that linear rules are learned more rapidly than more complicated rules and positive linear rules are learned more rapidly than negative linear rules.

It is well known in this literature (e.g., Brehmer, 1973) that the correlation coefficient is a composite measure that combines information about different aspects of a linear relationship. Most of the recent studies manipulate the value of r_{XY} by holding b_{YX} constant and varying $S^2_{Y\cdot X}$, although other manipulations (such as holding $S^2_{Y\cdot X}$ constant and varying b_{YX}) have been used. Although there

is some disagreement about the details, it appears that different aspects of performance are controlled by different features of the relationship. For example, Brehmer (1973) concluded that the final level of performance is determined by r_{XY} but that the rate of learning is determined by S^2_{YX}. Naylor and Domine (1981) suggested that S_Y/S_X may also play a role in determining final performance level and that rate of learning is influenced by some function of S_{YX} and S_X.

An important finding is that when the predictor and criterion have an imperfect linear relationship, subjects quickly learn to predict well, according to the slope measure discussed earlier. After an initial period of learning, the ratio b_{PX}/b_{YX} tends to hover close to a value of 1 under a variety of conditions (e.g., Brehmer, 1973). The slope ratio often, but not always (e.g., Naylor & Domine, 1981), tends to be larger for small values of r_{XY}.

However, performance does not become optimal, even after extended learning. The problem is that subjects do not consistently use the optimal linear prediction strategy that minimizes error. The consistency measure, r_{PX}, does not reach a value of 1 and is a function of r_{XY}. For example, Kuylenstierna and Brehmer (1981), using a value of .5 for r_{XY}, were unable to obtain values of r_{PX} much greater than .8, even when subjects were allowed to use memory aids (i.e., were allowed to plot running scatter diagrams as the predictor-criterion pairs were presented) and were given detailed instructions concerning the statistical nature of the task.

The lack of complete consistency is attributed to subjects' inability to deal with the inherent randomness or unpredictability of the task. Instead of inferring and using the optimal linear rule that minimizes prediction error, they engage in an extended period of hypothesis testing in an attempt to find a higher order rule that will predict perfectly. It is also possible that some of the prediction variability results from variability in the encoding of X, given that the values of the predictor variable are often given in terms of line lengths.

To the best of our knowledge, few process models have arisen from this literature. Brehmer (1974) discussed a hypothesis testing model that depicts the learning of inference tasks as a two-stage process. In the first stage subjects detect the appropriate rule, and in the second stage they learn to use it. In earlier papers (e.g., Brehmer, 1974), the rule detection process was characterized in terms of sampling hypotheses from a hierarchy of hypotheses. In more recent papers, verbal report data, as well as evidence that at least some subjects can learn complex rules, led Brehmer to suggest that subjects may construct hypotheses rather than sample them (e.g., Brehmer, 1980).

Our prediction task differed from the standard SPL task in a number of ways. We employed a cover story in an attempt to make the problems meaningful and presented subjects with only 30 prediction trials. Also, in the current study, once an X-Y pair had been presented, it remained on the screen. In the typical SPL study, once a prediction has been made and feedback presented, the X and Y information is no longer available for inspection. Nevertheless, prediction performance in the current study was roughly what might have been expected based on the SPL literature: good performance on the average but more than optimal variability in subjects' predictions.

We believe it would be unrealistic to have expected our subjects to use the optimal regression strategy consistently. They were not instructed to use a linear rule and were not told that the optimal strategy was to use a simple rule that minimized error even though this meant giving up any attempt to predict perfectly by finding some more complicated rule. Even if subjects knew they should be using a linear rule, the details of the rule would change as the subjects were presented with more stimulus pairs. Also, the spontaneous comments of some subjects suggested that they not only used information about the current value of X in predicting Y but superimposed various sequential strategies in an attempt to predict more accurately. For example, some subjects were more likely to produce larger predictions than warranted by the value of X on the current trial if their last few predictions had been smaller than the criterion value.

An advantage of the prediction data is that they provide a rich data base from which to begin developing and testing process models. One class of models might deal with how subjects construct and use different kinds of functional rules, in the spirit of the SPL literature. Functional rules may evolve from simple prediction strategies that are themselves not rules but encourage attention to aspects of the relationship that form the basis for rules. Such strategies may, for example, abstract regularities in the relationship between X and Y by reducing the distorting effects of error variance and thereby help to reveal the nature of the underlying relationship between X and Y. One such strategy would be (1) to divide values on the X variable into a number of equivalence classes, and (2) when given a value of X that falls into a particular class, to respond with the perceived "average" value for Y that has been registered for that class. In general, this strategy would result in predictions that had the same slope and intercept as the optimal regression line. The degree of variability in the predictions would be determined by the details of how classes of X are formed and changed as new information arrives and how impressions of average Y for a class are formed and updated. This model has similarities to prototype abstraction models in the concept learning literature.

An equally viable second class of models assumes that subjects base their predictions on previously seen exemplars, as proposed by exemplar models of concept learning. In this case, the value of X would serve as a probe of memory, and the prediction that would be generated would be either a Y that occurred with a previously seen X to which the probe was similar (Nosofsky, 1987) or possibly some amalgamation of the Ys recorded in the memory traces of similar Xs (Hintzman, 1986). Of course, as Estes (1986), Hintzman (1986), and Medin (1986) noted, it will be difficult to distinguish between these classes of models. Presumably, whether subjects per-

formed as though they used rules could be tested by giving them transfer conditions in which subjects were presented with values of X different from those they seen earlier.

We believe that future efforts should be directed away from the question of whether subjects can produce numerical judgments that reflect the correlation coefficient or some other measure of covariation and toward developing a better understanding of what information about the relationship between two variables people are sensitive to and how this information is used and represented. To this end, what people learn about event covariation and how they use what they have learned can probably best be investigated by developing process models and comparing the predictions of these models to the data of human subjects.

REFERENCES

ACHEN, C. H. (1982). *Interpreting and using regression*. Beverly Hills: Sage.

ALLOY, L. B., & TABACHNIK, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review*, **91**, 112-149.

BEACH, L. R., & SCOPP, T. S. (1966). Inferences about correlations. *Psychonomic Science*, **6**, 253-254.

BOBKO, P., & KARREN, R. (1979). The perception of Pearson product-moment correlations from bivariate scatterplots. *Personnel Psychology*, **32**, 313-325.

BREHMER, B. (1973). Single-cue probability learning as a function of the sign and magnitude of the correlation between cue and criterion. *Organizational Behavior & Human Performance*, **9**, 377-395.

BREHMER, B. (1974). Hypotheses about scaled relations in the learning of probabilistic learning tasks. *Organizational Behavior & Human Performance*, **11**, 1-27.

BREHMER, B. (1976). Transfer in single-cue probability learning. *Organizational Behavior & Human Performance*, **16**, 177-192.

BREHMER, B. (1980). Effects of cue validity on learning of complex rules in probabilistic inference tasks. *Acta Psychologica*, **44**, 201-210.

CLEVELAND, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, **74**, 829-836.

CLEVELAND, W. S., DIACONIS, P., & MCGILL, R. (1982). Variables on scatterplots look more highly correlated when the scales are increased. *Science*, **216**, 1138-1141.

CROCKER, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin*, **90**, 272-292.

ERLICK, D. E., & MILLS, R. G. (1967). Perceptual quantifications of conditional dependency. *Journal of Experimental Psychology*, **73**, 9-14.

ESTES, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.

HANUSHEK, E. A., & JACKSON, J. E. (1977). *Statistical methods for social scientists*. New York: Academic Press.

HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.

JENNINGS, D. L., AMABILE, T., & ROSS, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211-230). Cambridge, England: Cambridge University Press.

KUYLENSTIERNA, J., & BREHMER, B. (1981). Memory aids in the learning of probabilistic learning tasks. *Organizational Behavior & Human Performance*, **28**, 415-424.

LANE, D. M., ANDERSON, C. A., & KELLAM, K. L. (1985). Judging the relatedness of variables: The psychophysics of covariation detection. *Journal of Experimental Psychology: Human Perception & Performance*, **11**, 640-649.

MALMI, R. A. (1986). Intuitive covariation estimation. *Memory & Cognition*, **14**, 501-508.

MEDIN, D. L. (1986). Comment on "Memory storage and retrieval process in category learning." *Journal of Experimental Psychology: General*, **115**, 373-381.

NAYLOR, J. C., & DOMINE, R. K. (1981). Inferences based on uncertain data: Some experiments on the role of slope magnitude, instructions, and stimulus distribution shape on learning contingency relationship. *Organizational Behavior & Human Performance*, **27**, 1-31.

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

NOSOFSKY, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **13**, 87-108.

SLOVIC, P. (1974). Hypothesis testing in the learning of positive and negative functions. *Organizational Behavior & Human Performance*, **11**, 368-376.

SNIEZEK, J. A. (1986). The role of variable labels in cue probability learning tasks. *Organizational Behavior & Human Decision Processes*, **38**, 141-161.

STRAHAN, R. F., & HANSEN, C. J. (1978). Understanding correlation from scatterplots. *Applied Psychological Measurement*, **2**, 543-550.

WRIGHT, J. C., & MURPHY, G. L. (1984). The utility of theories in intuitive statistics: The robustness of theory-based judgments. *Journal of Experimental Psychology: General*, **113**, 301-322.

NOTES

1. There are also a number of studies that have presented relatively sophisticated subjects with scatter diagrams and asked them to estimate the correlation coefficient or in some cases degree of linear relationship (e.g., Bobko & Karren, 1979; Cleveland, Diaconis, & McGill, 1982; Strahan & Hansen, 1978). These studies are not considered here.

2. It should be noted that that an analogous concern can be expressed about binary variables. The delta coefficients or differences between conditional probabilities [which can be expressed as $a/(a+b) - c/(c+d)$ and $a/(a+c) - b/(b+d)$] correspond to the regression coefficients, and the phi coefficient corresponds to the correlation. The phi coefficient is the geometric mean of the two deltas, and all three measures will take on the same value only when the variances of X and Y are equal. These different measures of relationship are usually not distinguished, although it is conceivable that people are differentially sensitive to them.

APPENDIX A

A private consulting firm was hired by three companies to gather information about the relationship between the distance traveled by employees to their workplace and their work efficiency. For randomly selected groups of employees at each company, the consultants obtained two measures. The first measure, X, refers to the number of miles the employee travels to work each day, round trip. The second measure, Y, is work efficiency in percent of projects completed. The first set of data is for company A, the second for company B, and the third for company C.

X: Distance employee commutes round trip in miles per day

Y: Percent of projects completed

APPENDIX B

All subjects were presented with the first two paragraphs. The remainder differed for the observation and prediction groups:

In everyday life people routinely learn to make judgments about how strongly things are related. Some things are strongly related—for example, arm length and leg length—while others are not related at all—for example, length of index finger and intelligence. In this study, we are inves-

tigating how accurately people can judge the strength of such relationships.

You will begin by reading a brief passage that deals with whether the distance an employee travels to work is related to their efficiency on the job. By the end of this experiment you will have evaluated data from three companies. For each company the data will be presented as follows:

Observation Group

You will see two sets of 30 pairs of numbers where the first number indicates an employee's commuting distance and the second number indicates that employee's efficiency on the job (measured in percent projects completed). These numbers will give you some idea about the relationship between these two variables.

After you have gone through both sets of 30 pairs, all 60 pairs will be put on the screen so that you can look back at any of them you wish. You will then be asked to judge how strongly commuting distance and work efficiency are related for this company. You will indicate your judgment on a scale that goes from 0 (no relationship) to 100 (perfect relationship). Finally you will be asked how confident you are about your judgment. This procedure will be repeated for each of the three companies.

Prediction Group

You will see 30 pairs of numbers where the first number indicates an employee's commuting distance and the second number indicates that employee's efficiency on the job (measured in percent projects completed). These numbers will give you some idea about the relationship between these two variables.

Then you will be given only an employee's commuting distance and asked to predict, as accurately as you can, that employee's work efficiency. After you have made your prediction the actual value will be presented.

After you have gone through 30 pairs in this fashion, all 60 pairs will be put up on the screen so that you can look back at any of them you wish. You will then be asked to judge how strongly commuting distance and work efficiency are related for this company. You will indicate your judgment on a scale that goes from 0 (no relationship) to 100 (perfect relationship). Finally, you will be asked to indicate how confident you are about your judgment. This procedure will be repeated for each of the three companies.

(Manuscript received February 17, 1987;
revision accepted for publication November 23, 1987.)