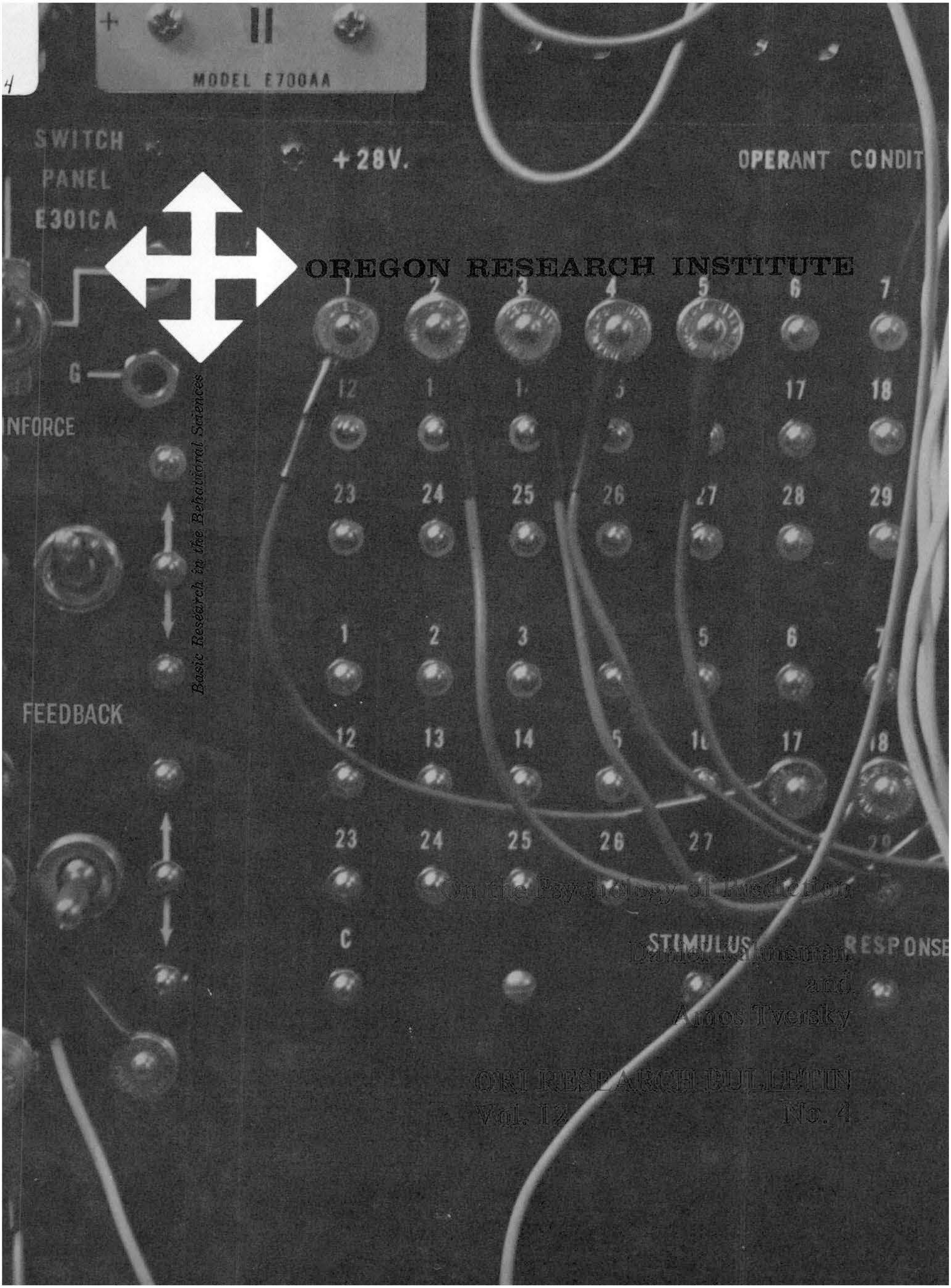


OREGON RESEARCH INSTITUTE

Basic Research in the Behavioral Sciences

On the Psychology of Prediction

Daniel Kahneman
and
Amos Tversky

ORI RESEARCH BULLETIN
Vol. 12 No. 4

OREGON RESEARCH INSTITUTE

On the Psychology of Prediction

Daniel Kahneman and Amos Tversky

The Hebrew University of Jerusalem

and

Oregon Research Institute

Research Bulletin
Vol. 12 No. 4

October 1972

Abstract

A judgmental heuristic--representativeness--is often applied in intuitive predictions. In predicting by representativeness, one selects or orders outcomes by the degree to which these outcomes represent one's impression of the case, with little or no regard for factors such as the prior probability of outcomes or predictive accuracy. The hypothesis that people predict by representativeness is supported in a series of studies, with both naive and sophisticated subjects. In nominal prediction, the ranking of outcomes by likelihood is shown to coincide with the ranking of outcomes by representativeness. Numerical predictions are shown to be non-regressive with respect to an evaluation of the impression on which they are based. In both cases, intuitive predictions violate the normative principles of prediction in fundamental ways.

ON THE PSYCHOLOGY OF PREDICTION [1]

Daniel Kahneman and Amos Tversky

The Hebrew University of Jerusalem

and

Oregon Research Institute

In this paper we explore the rules that determine intuitive predictions and judgments of confidence, and contrast these rules to the normative principles of statistical prediction. We are concerned with situations where the range of possible outcomes is specified in advance and one predicts or orders outcomes by their likelihood. Two classes of prediction will be discussed: category prediction and numerical prediction. In a categorical case, the prediction is given in nominal form; e.g., the winner in an election, the diagnosis of a patient, or a person's future occupation. In a numerical case, the predictions are given in numerical form; e.g., the future value of a particular stock, or of a student's grade-point average.

Three types of information are relevant to prediction: (i) prior or background information, (ii) specific evidence concerning the case about which a prediction is made, and (iii) expected accuracy of the prediction. A fundamental principle of the statistical theory of prediction is that expected accuracy controls the relative weights assigned to specific evidence and to prior information. When expected accuracy decreases, normative predictions become more regressive, i.e., closer to prior expectations.

In making intuitive judgments, people do not apply the logic of statistical prediction. Instead, we propose, they rely on a heuristic--representativeness--which sometimes yields reasonable answers and sometimes leads to severe and systematic errors. The essence of the representativeness heuristic is that judgments of likelihood are reduced to judgments of similarity (Kahneman & Tversky, 1972). In predicting by representativeness, one selects or orders outcomes (e.g., the professions among which an individual will choose) by the degree to which these outcomes represent one's impression of the case (e.g., of the personality of that individual). Thus, prediction by representativeness is determined by a comparison of the possible outcomes with the relevant input or impression. Considerations of prior probability of outcomes and expected accuracy do not enter into this comparison. To confirm the representativeness hypothesis, therefore, we show (i) that people predict the outcome which they see as most representative of the input information, and (ii) that they ignore factors that do not affect the representativeness relation, such as prior probability of outcomes, and expected predictive accuracy.

In the first section we investigate category predictions and show that they conform to an independent assessment of representativeness, or similarity, and that they are essentially independent of the prior probability of outcomes. In the next section we investigate numerical predictions and show that they are not properly regressive and are essentially unaffected by considerations of reliability. The source of fallacious intuitions concerning regression effects is discussed in the following section. An analysis which suggests that subjective confidence and predictive accuracy are often negatively correlated is presented in the final section.

CATEGORY PREDICTION

Likelihood, Similarity, and Base rate

Consider the following question, which was answered by 114 graduate students in psychology at three major universities in the U.S.:

Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

The preceding personality sketch of Tom W. was written during Tom's senior year in high school by a psychologist on the basis of projective tests. Tom W. is currently a graduate student. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W. is now a graduate student in that field. Write 1 next to the most probable choice, 2 next to the second most probable choice, etc. Ties are not allowed.

Insert Table 1 about here

The nine outcomes, i.e., alternative areas of graduate specialization are listed in Table 1. The first column of this table presents the mean of the ranks assigned to each outcome by the respondents who predicted Tom W.'s area of specialization.

A second group of 69 subjects[2] were given the description of Tom W., but were told neither his age at the time the description was written nor the information on which it was based. They were merely asked: "How similar is Tom W. to the typical graduate student in each of the following nine areas of graduate specialization? Please rank the nine areas etc." The second column in Table 1 presents the mean of the ranks assigned to each area by the subjects who evaluated Tom W.'s similarity to the typical graduate student in that area.

A third group of 65 subjects were given the list of nine areas of graduate specialization and were asked the following question: "Consider all first-year graduate students in the U.S. today. Please write down your best guesses

TABLE 1

Summary of Prediction and Similarity Data for Tom W., and

Estimated Base-Rates of the Nine Areas of Graduate Specialization

Outcomes	Mean likelihood rank	Mean similarity rank	Mean judged base rate
Business Administration	4.3	3.9	15%
Computer Sciences	2.5	2.1	7%
Engineering	2.6	2.9	9%
Humanities and Education	7.6	7.2	20%
Law	5.2	5.9	9%
Library Science	4.7	4.2	3%
Medicine	5.8	5.9	8%
Physical and Life Sciences	4.3	4.5	12%
Social Science and Social Work	8.0	8.2	17%

about the percentages of these students who are now enrolled in each of the following nine fields of specialization." The third column in Table 1 presents the mean of the estimated base rates for each of the nine areas.

The product-moment correlation between the mean judgments of likelihood and of similarity is .97, while the correlation between judged likelihood and estimated base-rate[3] is -.65. Evidently, judgments of likelihood essentially coincide with judgments of similarity and are quite unlike the estimates of base rates.

The likelihood judgments made by the psychology graduate students drastically violate the normative rules of prediction. More than 95% of these respondents judged that Tom W. is more likely to study computer science than humanities or education, although they were surely aware of the fact that there are many more graduate students in the latter field. According to the base-rate estimates shown in Table 1, the prior odds for humanities against computer science are about 3 to 1. (The actual odds are considerably higher.)

According to Bayes' rule it is of course possible to overcome the prior odds against Tom W. being in computer science rather than in humanities, if the description of his personality is both accurate and diagnostic. The graduate students in our study did not believe that these conditions were met. Following the prediction task, the respondents were asked to estimate the percentage of hits (i.e., correct first choices among the nine areas) which could be achieved with several types of information. The median estimate of hits was 23% for predictions based on projective tests, which compares to 53%, for example, for predictions based on high school seniors' reports of their interests and plans. Evidently, projective tests were held in low esteem. Nevertheless, the graduate students relied exclusively on information based on such tests and ignored the base-rate.

Immediately after they completed their predictions of Tom W.'s graduate career, the subjects encountered the following question.

> In fact, Tom W. is a graduate student in the School of Education and he is enrolled in a special program of training for the education of handicapped children. Please outline very briefly the theory which you consider most likely to explain the relation between Tom W.'s personality and his choice of career.

Our purpose in introducing this dissonance-inducing question was to find out whether sophisticated graduate students, skeptical about projective tests, would resolve the dilemma by simply denying the validity of the original personality description which was based on such tests. Tom's choice of special education as a profession is inconsistent with the expectations induced by the description of his personality. Indeed, Tom W. was seen as very dissimilar to the stereotype of graduate students in humanities and education and as very unlikely to specialize in this area. A sensible way to resolve the inconsistency is to deny the validity of the description, which was based on the flimsy evidence of projective tests, and to surmise that the "real" Tom W. never possessed the traits attributed to him in the description. However reasonable, this solution was not popular among the graduate students. Only 24% of the respondents even mentioned the possibility that the psychologist's description was inaccurate. The vast majority of the explanations of Tom's choice of career (including those given by the doubters) either focused on key features in the description such as Tom W.'s competence or his deep moral sense, or provided psychodynamic explanations in terms of threat avoidance or a need to interact with inferiors. In summary, sophisticated subjects not only relied uncritically on dubious information in making predictions, they also found it exceedingly difficult to discard their initial impression when it came in conflict with the facts.

An additional study in the same paradigm tests the hypothesis that a decrease in expected accuracy reduces confidence without affecting the pattern of predictions. The experimental material consisted of five thumbnail personality sketches of ninth-grade boys, allegedly written by a counselor on the basis of an interview. The design was the same as in the Tom W. study. For each description, subjects in one group (N=69) ranked the nine fields of graduate specialization (see Table 1) in terms of the similarity of the boy described to their "image of the typical first-year graduate student in that field." Following the similarity judgment, they estimated the base-rate frequency of the nine areas of graduate specialization. These estimates were shown in Table 1. The remaining subjects were told that the five descriptions were randomly selected from the population of first-year graduate students who had been interviewed when they were in the ninth grade. One group, the high-accuracy group (N=55) was told that "on the basis of such descriptions, students like yourself make correct predictions in about 55% of the cases." The low-accuracy group (N=50) was told that students' predictions in this task are correct in about 27% of the cases. For each description, the subjects ranked the nine fields according to "the likelihood that the person described is now a graduate student in that field." For each description they also estimated the probability that their first choice was correct.

The manipulation of expected accuracy had a significant effect on these probability judgments. The mean estimates were .70 and .56, respectively, for the high and low accuracy groups ($t = 3.72$, $p < .001$). However, the orderings of the nine outcomes produced under the low-accuracy instructions were not closer to the base-rate distribution than the orderings produced under the high-accuracy instructions. A product-moment correlation was computed for each judge, between the average rank he had assigned to each of the nine outcomes (over the five descriptions) and the base rate. This correlation is an overall measure of the degree to which the subject's predictions conform to the base-rate distribution. The average of these individual correlations was .13 for subjects in the high-accuracy group, and .16 for subjects in the low-accuracy group. The difference does not approach significance ($t = 0.42$, 103 df). This pattern of judgments violates the normative theory of prediction, according to which any decrease in expected accuracy should be accompanied by a shift of predictions toward the base rate.

Since the manipulation of accuracy had no effect on predictions, the two accuracy groups were pooled. The following analyses were the same as in the Tom W. study. For each description, two correlations were computed: (i) between mean likelihood judgment and mean similarity judgment; and (ii) between mean likelihood judgment and mean base rate. These correlations are shown in Table 2, with the outcome judged most likely for each description. The correlations between prediction and similarity are consistently

---

Insert Table 2 about here

---

high. In contrast, there was no systematic relation between prediction and base rate: the correlations vary widely depending on whether the most representative outcomes for each description happen to be frequent or rare.

Here again, considerations of base rate were neglected. In the statistical theory, one is allowed to ignore the base rate only when one expects to be infallible. In all other cases, an appropriate compromise must be found between the ordering suggested by the description and the ordering of the base rates. It is hardly believable that a cursory description of a fourteen-

TABLE 2

Product Moment Correlations of Mean Likelihood Ranks

with Mean Similarity Ranks and with Base-rate

Description #	1	2	3	4	5
Modal first prediction:	Law	Computer Science	Medicine	Library Science	Business Administration
Correlation with mean similarity ranks	.93	.96	.92	.88	.88
Correlation with base rate	.33	-.35	.27	-.03	.62

year-old child based on a single interview could justify the degree of infallibility implied by the predictions of our subjects.

Following the five personality descriptions, the subjects were given an additional problem:

> About Don you will be told nothing except that he participated in the original study and is now a first-year graduate student. Please indicate your ordering and report your confidence for this case as well.
>
> Please rank the following nine fields in terms of how likely it is that Don is now a graduate student in that field.

For Don the correlation between the likelihood rank and the base rate was .74. Thus, the knowledge of base rates, which was not applied when a description was given, was utilized when no specific information was available.

Prior vs. Individuating Information

The next study provides a more stringent test of the hypothesis that intuitive predictions are dominated by specific or individuating information with little or no regard for prior probabilities. In this study, the individuating information was scanty, and the prior probabilities were made exceptionally salient, and compatible with the response mode.

Subjects were presented with the following cover story.

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written.
>
> You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.
>
> The same task has been performed by a panel of experts, who were highly accurate in assigning probabilities to the various descriptions. You will be paid a bonus to the extent that your estimates come close to those of the expert panel.

These instructions were given to a group of 85 subjects (the low engineer, or L group). Another group of 86 subjects (the high engineer, or H group) were given identical instructions except for the prior probabilities: they were told that the sample from which the descriptions had been drawn consisted of 70 engineers and 30 lawyers. All subjects were presented with the same five descriptions. One of the descriptions follows:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful, and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.
>
> The probability that Jack is one of the 30 engineers in the sample of 100 is ______%

Following the five descriptions, the subjects encountered the null description:

> Suppose now that you are given no information whatsoever about an individual chosen at random from the sample.
>
> The probability that this man is one of the 30 engineers in the sample of 100 is _____%

In each of the two groups, half the subjects were asked to evaluate the probability that each description belonged to an engineer (as in the example above), while the other subjects evaluated the probability that each description belonged to a lawyer. This manipulation had no effect. The median probabilities assigned to the outcomes engineer and lawyer in the two different forms added to about 100% for all descriptions. Consequently, the analysis and discussion of the results are stated in terms of the outcome engineer.

The design of this experiment permits the calculation of the normatively appropriate pattern of judgments. The derivation relies on Bayes' formula, in odds form. Let O denote the odds that a particular description belongs to an engineer rather than to a lawyer. According to Bayes' rule

$$O = QR$$

where Q denotes the prior odds that the randomly selected description belongs to an engineer rather than to a lawyer; and R is the likelihood ratio for a particular description, i.e., the ratio of the probability that a person randomly drawn from a population of engineers will be so described to the probability that a person randomly drawn from the population of lawyers will be so described.

For the H group, who were told that the sample consists of 70 engineers and 30 lawyers, the prior odds Q_H equals 70/30. For the L group , the prior odds Q_L equals 30/70. Thus, for each description, the ratio of the posterior odds for groups H and L equals:

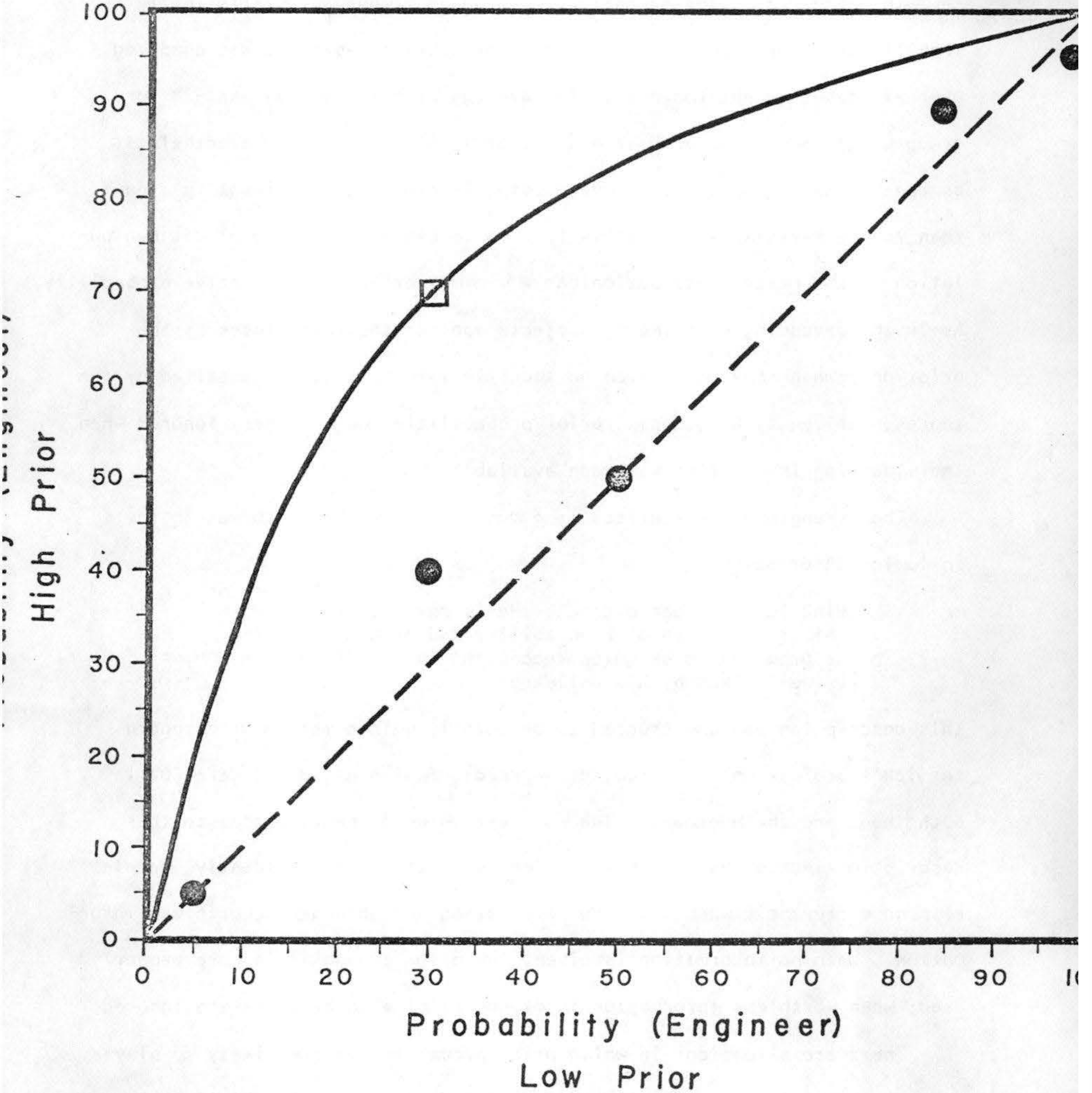
$$\frac{O_H}{O_L} = \frac{Q_H R}{Q_L R} = \frac{Q_H}{Q_L} = \frac{7/3}{3/7} = 5.44 .$$

Since the likelihood ratio is cancelled in this formula, the same value of O_H/O_L should obtain for all descriptions. In the present design, therefore, the correct effect of the manipulation of prior odds can be computed without knowledge of the likelihood ratio.

Insert Figure 1 about here

Figure 1 presents the median probability estimates for each description, under the two conditions of prior odds. For each description, the median estimate of probability when the prior is high (Q_H = 70/30) is plotted against the median estimate when the prior is low (Q_L = 30/70). According to the normative equation developed in the preceding paragraph, all points should lie on the curved (Bayesian) line. In fact, only the empty square which corresponds to the null description falls on this line: when given no description, subjects judged the probability to be 70% under Q_H and 30% under Q_L. In the

Figure 1. Median judged probability (Engineer) for five descriptions and for the null description (square symbol) under high and low prior probabilities. The curved line represents the normative Bayesian curve.



other five cases the points fall close to the identity line.

The effect of prior probability, although slight, is statistically significant. For each subject the mean probability estimate was computed over all cases except the null. The average of these values was 50% for group L and 55% for group H ($t = 3.23$, $df = 169$, $p < .01$). Nevertheless, as can be seen from Figure 1, every point is closer to the identity line than to the Bayesian line. It is fair to conclude that the explicit manipulation of the prior distribution had a minimal effect on subjective probability. As in the preceding experiment, subjects applied their knowledge of the prior only when they were given no specific information. As entailed by the representativeness hypothesis, prior probabilities were largely ignored when individuating information was made available.

The strength of this effect is demonstrated by the responses to the following description:

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

This description was constructed to be totally uninformative with regard to Dick's profession. Our subjects agreed: median estimates were 50% in both the L and the H groups. The contrast between the responses to this description and to the null description is instructive. Evidently, people respond differently when given no information and when given worthless information. When no information is given, the prior probabilities are properly used; when worthless information is given--prior probabilities are ignored.

There are situations in which prior probabilities are likely to play a more substantial role. In all the examples discussed so far distinct stereotypes were associated with the alternative outcomes, and judgments were controlled, we suggest, by the degree to which the descriptions appeared representative of these stereotypes. In other problems, the outcomes are more naturally viewed as segments of a dimension. Suppose, for example, that one is asked to judge the probability that each of several students will receive a fellowship. In this problem, there are no well-delineated stereotypes of recipients and non-recipients of fellowships. Rather, it is natural to regard the outcome (i.e., the reception of a fellowship) as determined by a cut-off point along the dimension of academic achievement or ability. Prior probabilities, i.e., the percentage of recipients of fellowships in the relevant group, are used to define the outcomes by locating the cut-off point. Consequently, they are not likely to be ignored. In addition, we would expect some effects of of extreme prior probabilities even in the presence of clear stereotypes of the outcomes. The delineation of the conditions under which prior information is used or discarded deserves further investigation.

One of the fundamental principles of the normative theory of prediction is that prior probability, which summarizes what we knew about the problem before receiving any specific evidence, remains relevant even after such evidence is obtained. Bayes' rule translates this qualitative principle into a multiplicative relation between prior odds and the likelihood ratio. Our subjects, however, failed to integrate prior probability with specific evidence. When exposed to a description--however scanty or suspect--of Tom W., or of Dick (the engineer/lawyer), they apparently felt that the distribution of occupations in his group was no longer relevant. The failure to appreciate the relevance of prior probability in the presence of specific evidence is perhaps one of the most significant departures of intuition from the normative theory of prediction.

The discarding of prior probabilies is not limited to the evaluation of psychological evidence. As an illustration, consider the following problem:

> Two cab companies, the Blue and the Green, operate in a given city. 85% of the cabs in the city are Blue and the remaining 15% are Green. A cab was involved in a hit-and-run accident at night.

One group of subjects (N=36) were asked to estimate (in percentages) the likelihood that a Blue cab rather than a Green cab was involved in the accident. As expected, the observed estimates centered around the base-rate percentage. The median estimate was 80%.

A second group of subjects (N=35) were given additional information as follows.

> A witness identified the cab as a Green cab. The court tested his ability to distinguish a Blue cab from a Green cab at night, and found that he was able to make correct identifications in 4 out of 5 tries.

Again, the subjects were asked to assess the likelihood that a Blue cab rather than a Green cab was involved in the accident. According to Bayes' rule, the estimated value should be slightly higher than 50% because the prior odds (which points to a Blue cab) are slightly higher than the likelihood ratio based on the testimony of a fallible witness (which points to a Green cab). If, on the other hand, the prior probability is completely suppressed by the witness' testimony, then the judged likelihood that a Blue cab rather than a Green cab was involved will be 20%. This was precisely the median estimate of our respondents. Much as we would like to, we have no reason to believe that the typical juror does not evaluate evidence in this fashion.

NUMERICAL PREDICTION

A fundamental rule of the normative theory of prediction is that the variability of predictions, over a set of cases, should increase with predictive accuracy. This rule applies to both category predictions and numerical predictions. If the achievable correlation between predictions and outcomes is zero, the variance of predictions should also be zero, i.e., the same prediction should be made in all cases. In category prediction, one would always predict the most frequent category. In numerical prediction, one would predict the mean, the mode, the median, or some other value depending on the loss function. As predictive accuracy increases, one should gradually extend the range of predictions. When predictive accuracy is perfect, and only then, should the variance of predictions match the variance of the criterion.

Imperfect predictive accuracy can result from unreliability in the measurement of the input and outcome variables, or from an imperfect underlying relation between them. As far as normative prediction is concerned, however, it is immaterial whether predictive accuracy is limited by unreliability or by imperfect validity. In either case, predictions should be regressive, and the variance of predictions should be smaller than the expected variance of outcomes.

In contrast, prediction by representativeness will typically be insufficiently regressive, for two reasons. First, people tend to view both the input information and the outcome as highly representative of the same underlying process, and therefore expect them to be more similar or consistent than they actually are. Indeed, in many situations people tend to view variables which are merely related as equivalent manifestations of the same trait. In those situations, as will be shown below, the pattern of predictions is completely nonregressive. Second, if the judge predicts the outcome which best represents the input information, then his predictions will be unaffected by his knowledge that this information is not reliable. Contrary to the normative model, predictions by representativeness will be no more regressive when reliability is low than when it is high.

These implications of the representativeness heuristic are tested in the following studies, in which subjects were given various types of information concerning hypothetical college students and were required to predict their grade-point average (GPA). The GPA was chosen as an outcome variable, because its correlates and distributional properties are well known to the subject population.

Predictions from Percentile Scores

Three groups of subjects participated in the experiment. Subjects in all groups predicted the GPA of ten hypothetical students, on the basis of a single percentile score obtained by each of these students. The same set of percentile scores was presented to all groups, but the three groups received different interpretations of the input variable, as follows.

1. Percentile GPA. The subjects in Group 1 (N=32) were told that "for each of several students you will be given a percentile score representing his academic achievements in the freshman year, and you will be asked to give your best guess about his GPA for that year." It was explained to the subject that "a percentile score of 65, for example, means that the GPA achieved by this student is better than that achieved by 65% of his class, etc."

2. Mental Concentration. The subjects in Group 2 (N=37) were told that "the test of mental concentration measures one's ability to concentrate and to extract all the information conveyed by complex messages. It was found that students with high grade-point averages tend to score high on the mental concentration test and vice versa. However, performance on the mental concentration test was found to depend on the mood and mental state of the person at the time he took the test. Thus, when tested repeatedly, the same person could obtain quite different scores, depending on the amount of sleep he had the night before or how well he felt that day."

3. Sense of humor. The subjects in Group 3 (N=35) were told that "the test of sense of humor measures the ability of people to invent witty captions for cartoons and to appreciate humor in various forms. It was found that students who score high on this test tend, by-and-large, to obtain a higher GPA than students who score low. However, it is not possible to predict GPA from sense of humor with high accuracy."

In the present design all subjects predicted GPA on the basis of the same set of percentile scores. Group 1, who predicted GPA from percentile GPA, merely expressed the given scores in different units. Groups 2 and 3, on the other hand, predicted GPA from rather remote variables. Group 2 predicted from a potentially valid but unreliable test of mental concentration, while Group 3 predicted from sense of humor, which is not generally viewed as closely related to academic ability. Normative considerations, therefore, dictate that the predictions of GPA from mental concentration and sense of humor should be more regressive than the predictions of GPA from percentile GPA.

The representativeness hypothesis suggests that mental concentration is likely to be viewed as a measure of academic achievement, and that the unreliability of that test will be ignored. Consequently, predictions from the unreliable test of mental concentration were expected to be no more regressive than predictions from percentile GPA. Group 3 was introduced as a control, to anticipate the objection that non-regressive predictions in Group 2 could be due to subjects ignoring the labels of the input variables.

The mean predictions assigned to the 10 percentile scores by the three groups are shown in Figure 2. It is evident in the figure the the predictions of Group 2 are no more regressive than the predictions of Group 1, while the predictions of Group 3 appear more regressive.

Insert Figure 2 about here

Four indices were computed within the data of each individual subject: the mean of his predictions, the standard deviation of his predictions, the slope of the regression of predicted GPA on the input scores, and the product-moment correlation between them. The mean of these values for the three groups are shown in Table 3.

Insert Table 3 about here

It is apparent in the table that the subjects in all three groups produced orderly data, as evinced by the high correlations between inputs and predictions (the average correlations were obtained by transforming individual values to Fisher's z). The results of planned comparisons between groups 1 and 2 and between groups 2 and 3 confirm the pattern observed in Figure 2. There are no significant differences between the predictions from percentile GPA and from mental concentration. Thus, people fail to regress when predicting from an input--however unreliable-- which is regarded as a measure of the outcome variable.

The predictions from sense of humor, on the other hand, are regressive, although not enough. The correlation between GPA and sense of humor inferred from a comparison of the regression lines is above .70. In addition, the predictions from sense of humor are significantly higher than the predictions from mental concentration. There is also a tendency for predictions from mental concentration to be higher than predictions based on percentile GPA. These findings are typical. When predicting the academic achievement of

Figure 2. Predictions of GPA from percentile scores on three variables.

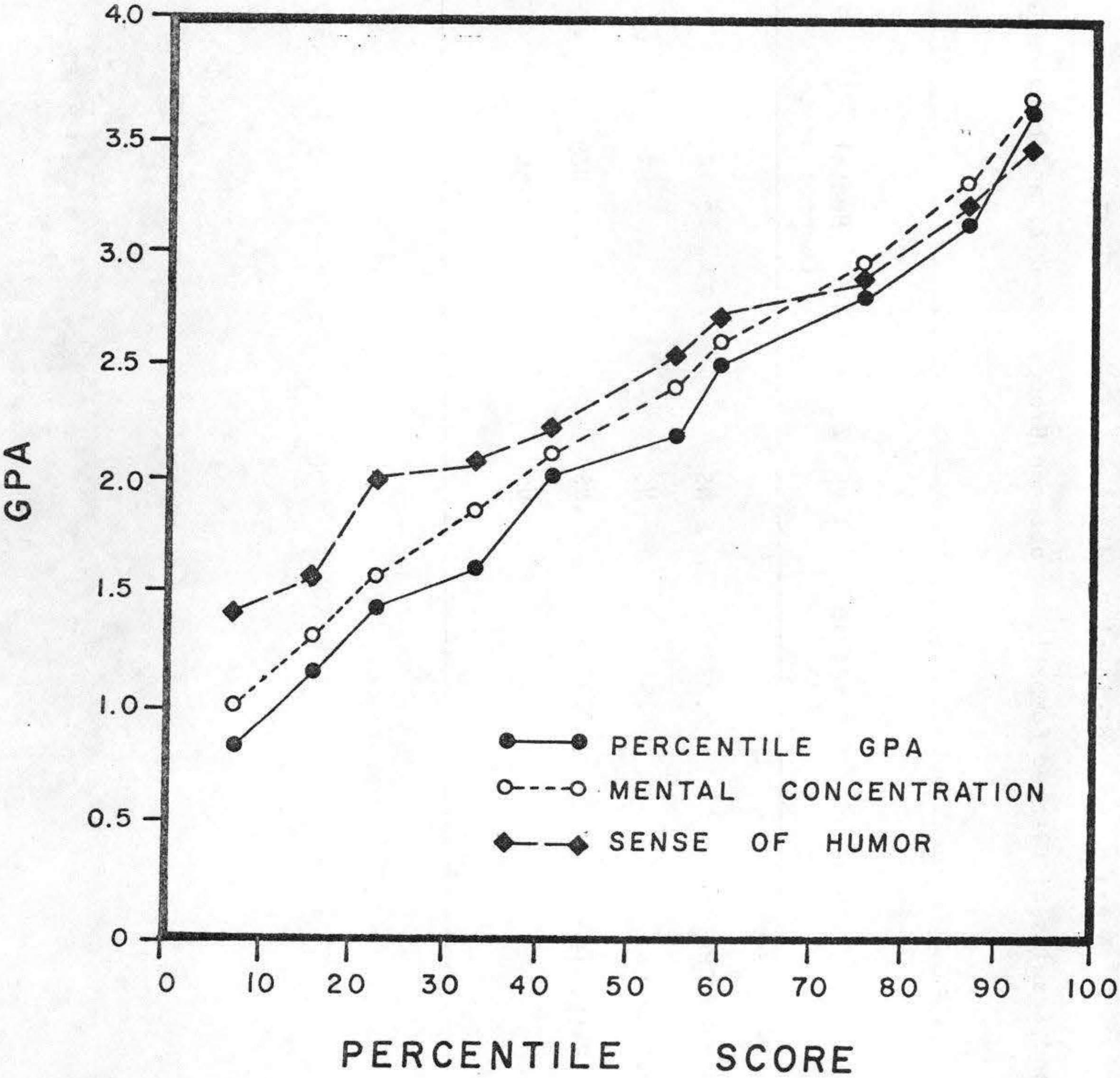


TABLE 3

Averages of Individual Prediction Statistics for Three Input Variables

and Results of Planned Comparisons between Groups 1 and 2, and between Groups 2 and 3

Statistic	Group 1. Percentile GPA	1 vs. 2	2. Mental Concentration	2 vs. 3	3. Sense of Humor
Mean predicted GPA	2.27	NS	2.35	.05	2.46
SD of predictions	.91	NS	.87	.01	.69
Slope of regression	.030	NS	.029	.01	.022
Correlation	.97	NS	.95	NS	.94

an individual on the basis of imperfect information, subjects tend to give him the benefit of the doubt, i.e., they respond to a reduction of validity by raising the predicted level of performance.

Predictions from Descriptions

Suppose one is told that a college freshman has been described by a counselor as intelligent, self-confident, well-read, hard-working, inquisitive. Consider the two following questions that might be asked about this description: 1. How does this student impress you with respect to his academic ability? What percentage of freshmen do you believe would impress you more? 2. What is your estimate of the GPA that this student will obtain? What is the percentage of freshmen who obtain a higher GPA?

There is a subtle but important difference between the two questions. In the first, you evaluate your own impression; in the second, you predict an outcome. Since there is surely greater uncertainty about the answer to question 2 than about the answer to question 1, your prediction should be more regressive than your evaluation. That is, the percentage you give as an answer to question 2 should be closer to 50% than the percentage you give as an answer to question 1. To observe the difference between the two questions, consider the possibility that the description is unreliable. This information should have no effect on your answer to question 1: the ordering of descriptions with respect to the impressions they make on you is independent of their reliability. In answering question 2, on the other hand, you should be regressive to the extent that you suspect the description to be unreliable or your prediction to be valid.

The essence of the representativeness hypothesis, however, is that people predict by evaluation. That is, they answer question 2 as if it were question 1. Several studies were conducted to test this hypothesis.

In each of these studies the subjects were given descriptive information concerning a set of cases. An evaluation group evaluated the quality of each description relative to a stated population, and a prediction group predicted future performance. The judgments of the two groups were compared to test whether predictions are more regressive than evaluations.

In two studies, subjects were given descriptions of college freshmen, alleged written by a counselor on the basis of an interview administered to the entering class. In the first study, each description consisted of five adjectives, referring to intellectual qualities and to character, as in the example cited above. In the second study, the descriptions were paragraph-length reports, including details of the student's background and of his current adjustment to college. In both studies the evaluation groups were asked to evaluate each one of the descriptions by estimating "the percentage of students in the entire class whose descriptions indicate a higher academic ability." The prediction groups were given the same descriptions and were asked to predict the GPA achieved by each student at the end of his freshman year, and his class standing in percentiles.

The results of both studies are shown in Figure 3, which plots, for each description, the mean prediction of percentile GPA against the mean evaluation. The only systematic discrepancy between predictions and evaluations is observed in the adjective study where predictions were consistently higher than the corresponding evaluations. The standard deviation of predictions or evaluations was computed within the data of each subject. A comparison of these values indicated no significant differences in variability between the evaluation and the prediction groups, within the range of values under study. In the adjectives study, the average standard deviation was 25.7 for the evaluation group (N=38) and 24.0 for the prediction group (N=36); (t = 1.25, 72 df, NS). In the reports study, the average standard deviation was 22.2 for the evaluation group (N=37) and 21.4 for the prediction group (N=63); (t = 0.75, 98 df, NS). In both studies the prediction and the evaluation groups produced equally extreme judgments, although the former predicted a remote objective criterion on the basis of sketchy interview information, while the latter merely evaluated the subjective impression obtained from each description. In the statistical theory of prediction, the observed equivalence between prediction and evaluation could be justified only if predictive accuracy were perfect, a condition which could not conceivably be met in these studies.

Insert Figure 3 about here

Further evidence for the equivalence of evaluation and prediction was obtained in a master's thesis by Beyth (1972). She presented three groups of subjects with seven paragraphs, each describing the performance of a student teacher during a particular practice lesson. The subjects were students in a statistics course at the Hebrew University. They were told that the descriptions had been drawn from among the files of 100 elementary school teachers who, five years earlier, had completed their teacher training program. Subjects in an evaluation group were asked to evaluate in percentile scores relative to the stated population the quality of the lesson described in the paragraph. Subjects in a prediction group were asked to predict in percentile scores the current standing of each teacher, i.e., his overall teaching competence five years after the description was written. An evaluation-prediction group performed both tasks. As in the studies described above, the differences between evaluation and prediction were not significant. This result held in both the between-subject and within-subject comparisons. Although the judges were undoubtedly aware of the multitude of factors that intervene between a single trial lesson and teaching competence five years later, this knowledge did not cause their predictions to be more regressive than their evaluations.

Figure 3A. Predicted percentile GPA as a function of percentile evaluation, for adjectives.

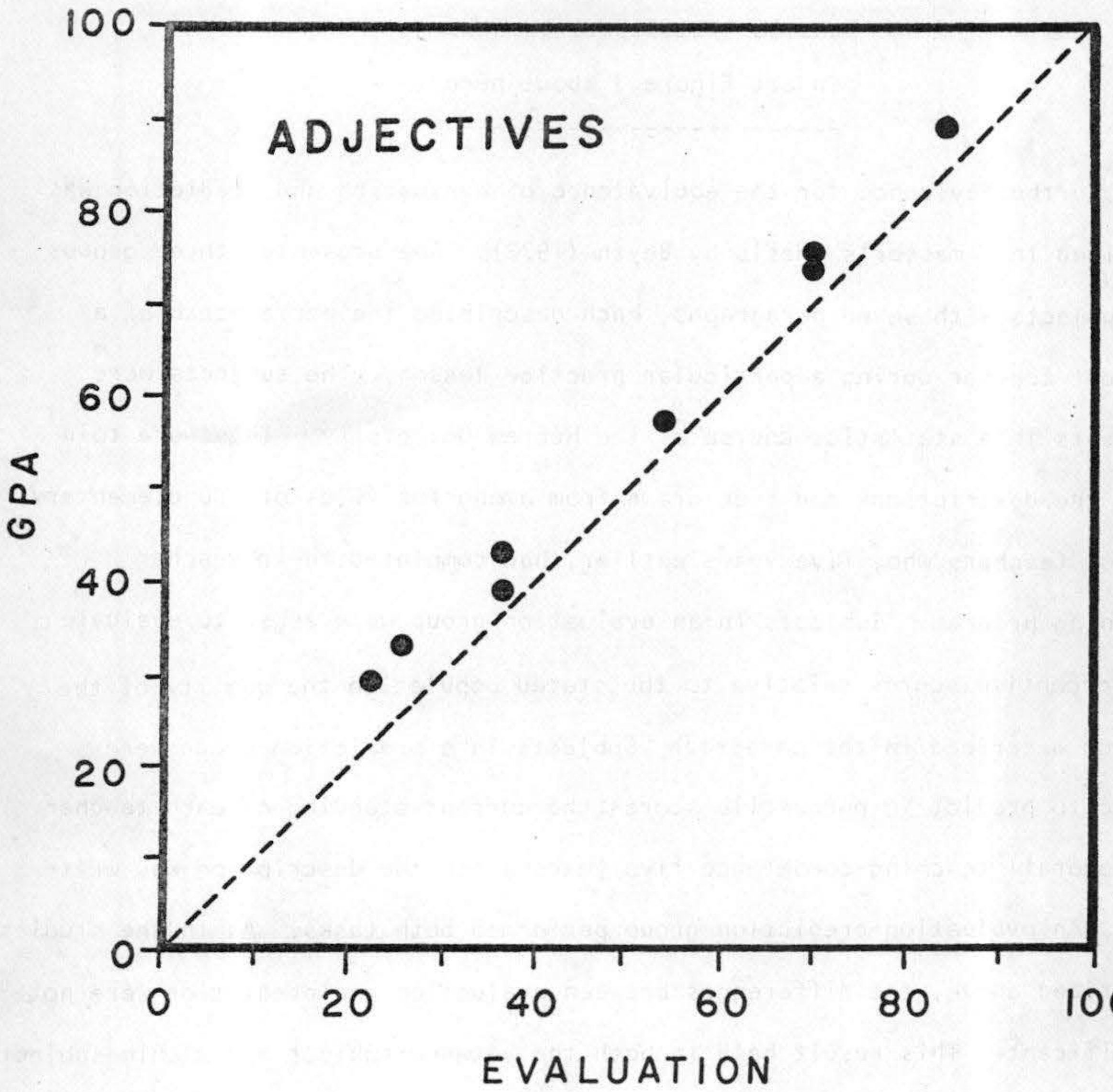
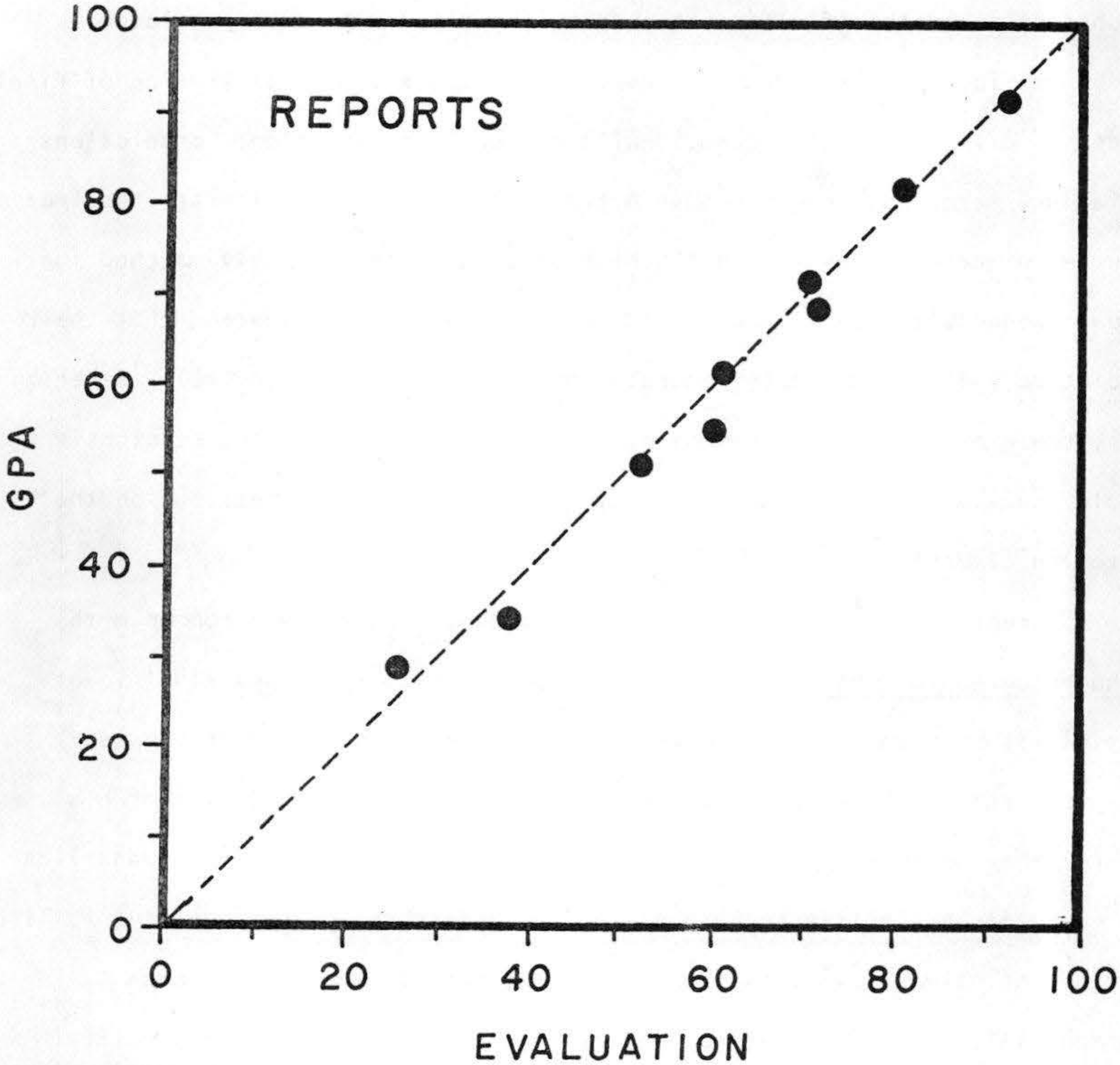


Figure 3B. Predicted percentile GPA as a function of percentile evaluation, for reports.



Prediction-by-evaluation has also been observed in a real-life setting with highly experienced judges in the Officer Selection Board of the Israel Army. The officers who participate in the assessment team normally evaluate candidates on a seven-point scale at the completion of several days of testing and observation. For the purposes of the study, they were required in addition to predict, for each successful candidate, the final grade that he would obtain in officer training school. In over two hundred cases, assessed by a substantial number of different judges, the distribution of predicted grades was found to be virtually identical to the actual distribution of final grades in officer training school, with one obvious exception: predictions of failure were less frequent than actual failures. In particular, the frequencies of predictions in the two highest categories precisely matched the actual frequencies. All judges were aware of research indicating that their predictive validity was only moderate (on the order of .20 to .40). Nevertheless, their predictions were non-regressive, in accord with the hypothesis that people predict by evaluation, i.e., by projecting their impressions on the criterion dimension.

Apparently, both naive and experienced observers share a common myth --<u>the holographic model of personality</u>--according to which any significant particle of behavior, however unreliable or remote, represents the essential features of the actor's personality and allows the prediction of his future behavior in a wide variety of settings. This belief, which underlies the elaborate projective techniques for the assessment of personality, is also part of the naive intuitions about the structure of behavior (Chapman & Chapman, 1967). The exaggerated belief in the consistency of personality structures and the expectation that measures which appear to represent the same underlying trait will be highly correlated may explain why layment and psychologists like are perennially surprised by the low correlations between seemingly interchangeable measures of honesty, of risk taking, of aggression, and of dependency (Mischel, 1968).

Methodological Considerations

The preceding section has reported several tests of the representativeness hypothesis. These tests showed that predictions from an unreliable variable (e.g., mental concentration) were no more regressive than judgments which merely required a change of scale (e.g., from percentile GPA to GPA). They also showed that predictions are no more regressive than the evaluations of corresponding impressions. It should be emphasized that the representativeness hypothesis implies that some judgments (e.g., predictions) will be no more regressive than other judgments (e.g., evaluations). It does not imply that either of these judgments will be non-regressive relative to the input scores on which they are based. The hypothesis that people predict by evaluation cannot be tested, therefore, by a direct comparison of judgments to inputs.

The present set of studies was concerned with situations in which people make predictions on the basis of information that is available to them prior to the experiment in the form of stereotypes (e.g., of an engineer) and expectations concerning relationships between variables. Outcome feedback was not provided, and the number of judgments required of each subject was small. In contrast, most studies of numerical prediction have been concerned with the learning of functional or statistical relations among variables. These studies typically involve a large number of trials and various forms of outcome feedback. (Some of this literature has been reviewed by Slovic & Lichtenstein, 1971.) In studies of repetitive predictions with feedback, subjects appear to predict by selecting outcomes so that the entire sequence or pattern of predictions is highly representative of the distribution outcomes. For example, subjects in probability learning studies generate sequences of predictions which roughly match the statistical characteristics of the sequence of outcomes. Similarly, subjects in numerical prediction tasks approximately reproduce the scatterplot, i.e., the joint distribution of inputs and outcomes (see e.g., Gray, 1968). To do so, people resort to a mixed strategy; for any given input, they generate a distribution of different predictions. These predictions represent the fact that any one input is followed by different outcomes on different trials. Evidently, the rules of prediction are different in the two paradigms, although representativeness is involved in both. In the feedback paradigm, subjects produce response sequences which represent the entire pattern of association between inputs and outcomes. In the situations explored in the present paper, subjects select the prediction which best represents their impressions of each individual case. The two approaches lead to different violations of the normative rules: the representation of uncertainty through a mixed strategy in the feedback paradigm, and the discarding of uncertainty through prediction-by-evaluation in the present paradigm.

UNDERSTANDING REGRESSION

Regression effects are all about us. In our experience, most outstanding fathers have somewhat disappointing sons, brilliant wives have duller husbands, the ill-adjusted tend to adjust and the fortunate are eventually stricken by ill luck. In spite of these encounters, people do not acquire a proper notion of regression. First, they do not expect regression in many situations where it is bound to occur. Second, as any teacher of statistics will attest, a proper notion of regression is extremely difficult to acquire. Third, when people observe regression, they typically invent spurious dynamic explanations for it.

What is it that makes the concept of regression counterintuitive and difficult to acquire and apply? We suggest that the major source of the difficulty is that regression effects typically violate the intuition that the predicted outcome should be maximally representative of the input information.

To illustrate the persistence of non-regressive intuitions despite considerable exposure to statistics, we presented the following problem to our sample of graduate students in psychology.

> A problem of testing. A randomly selected individual has obtained a score of 140 on a standard IQ test. Suppose that an IQ score is the sum of a "true" score and a random error of measurement which is normally distributed.
>
> Please give your best guess about the 95% upper and lower confidence bounds for the true IQ of this person. That is, give a high estimate such that you are 95% sure that the true IQ score is, in fact, lower than that estimate, and a low estimate such that you are 95% sure that the true score is in fact higher.

In this problem, the respondents were told to regard the observed score as the sum of a "true" score and an error component. Since the observed score is considerably higher than the population mean, it is more likely than not that the error component is positive and that this individual will obtain a somewhat lower score on subsequent tests. The majority of subjects (73 of 108), however, stated confidence intervals that were symmetric around 140, failing to express any expectation of regression. Of the remaining 35 subjects, 24 stated regressive confidence intervals and 11 stated counter-regressive intervals. Thus, most subjects ignored the effects of unreliability in the input, and predicted as if the value of 140 was a true score. The tendency to predict as if the input information were error-free has been observed repeatedly in this paper.

The occurrence of regression is sometimes recognized, either because we discover regression effects in our own observations, or because we are explicitly told that regression has occurred. When recognized, a regression effect is typically regarded as a systematic change that requires substantive explanation. Indeed, many spurious explanations of regression effects have been offered in the social sciences.[4] Dynamic principles have been invoked to explain why businesses which did exceptionally well at one point in time tend to deteriorate subsequently, and why training in interpreting facial expressions is beneficial to trainees who scored poorly on the pretest and detrimental to those who did best. Some of these explanations might not have been offered, had the authors realized that given two variables of equal variance the following two statements are logically equivalent: (i) Y is regressive with respect to X; (ii) the correlation between Y and X is less than unity. Explaining regression, therefore, is tantamount to explaining why a correlation is less than unity.

As a final illustration of how difficult it is to recognize and properly interpret regression, consider the following question which was put to our sample of graduate students. The problem described actually arose in the experience of one of the authors.

> A problem of training. The instructors in a flight school adopted a policy of consistent positive reinforcement recommended by psychologists. They verbally reinforced each successful execution of a flight maneuver. After some experience with this training approach, the instructors claimed that, contrary to psychological doctrine, high praise for good execution of complex maneuvers typically results in a decrement of performance on the next try.
>
> What should the psychologist say in response?

Regression is inevitable in flight maneuvers because performance is not perfectly reliable and progress between successive maneuvers is slow. Hence, pilots who did exceptionally well on one trial are likely to deteriorate on the next, regardless of the instructors' reaction to the initial success. The experienced flight instructors actually discovered the regression but attributed it to the detrimental effect of positive reinforcement. This true story illustrates a saddening aspect of the human condition. We normally reinforce others when their behavior is good and punish them when their behavior is bad. By regression alone, therefore, they are most likely to improve after being punished and most likely to deteriorate after after being rewarded. Consequently, we are exposed to a lifetime schedule in which we are most often rewarded for punishing others, and punished for rewarding.

Not one of the graduate students who answered this question mentioned regression in their discussion of the problem. Instead, they proposed that verbal reinforcements might be ineffective for pilots, or that they could lead to overconfidence. Some students even doubted the validity of the instructors' impressions and dicussed possible sources of bias in their perception of the situation. These respondents had undoubtedly been exposed to a thorough treatment of statistical regression. Nevertheless, they failed to recognize an instance of regression when it was not couched in the familiar terms of the height of fathers and sons. Evidently, statistical training alone does not change fundamental intuitions about uncertainty. The same conclusion was reached in a study of intuitions of sophisticated research psychologists concerning the significance and replicability of research results (Tversky & Kahneman, 1971).

THE ILLUSION OF VALIDITY

We have proposed that people predict by selecting the outcome that is most representative of the input. We further proposed that the degree of confidence that a person has in his prediction reflects the degree to which the selected outcome is more representative of the input information than are other outcomes.

A major determinant of representativeness in the context of numerical prediction with multiattribute inputs (e.g., score profiles) is the consistency, or the coherence, of the input. The more consistent the input, the more representative the predicted score will appear, and the greater the confidence in that prediction. For example, people predict an overall B average with more confidence on the basis of two B grades in two separate introductory classes, than on the basis of an A and a C. Indeed, internal variablility or inconsistency of the input has been found to decrease confidence in predictions (Slovic, 1966).

The intuition that consistent profiles allow greater predictability than inconsistent profiles is compelling. It is worth noting, however, that this belief is incompatible with the commonly applied multivariate models of prediction (e.g., the normal linear model) in which expected predictive accuracy is independent of within-profile variability.

Consistent profiles will typically be encountered when the judge predicts from highly correlated scores. Inconsistent profiles, on the other hand, are more frequent when the intercorrelations are low. Because confidence increases with consistency, confidence will generally be high when the input variables are highly correlated. However, given input variables of stated validity, the multiple correlation with the criterion is inversely related to the correlations among the inputs. Thus, a paradoxical situation arises where high intercorrelations among inputs increase confidence and decrease validity. We have observed this paradoxical effect in several studies in which the validity of the separate input scores was held constant while the correlation between the inputs was manipulated. Subjects were invariably more confident in their predictions when the correlation between the inputs was high than when it was low. That is, they were more confident in a context of inferior predictive validity.

Another finding observed in many prediction studies, including our own, is that confidence is a J-shaped function of the predicted level of performance (see e.g., Johnson, 1972). Subjects predict outstandingly high achievement with very high confidence, and utter failure with more confidence than mediocre performance. As we saw earlier, intuitive predictions are often insufficiently regressive. The discrepancies between predictions and outcomes, therefore, are largest at the extremes. The J-shaped confidence function entails that subjects are most confident in predictions that are most likely to be off the mark.

The foregoing discussion suggests that confidence in predictions is determined primarily by specified characteristics of the impression one has formed, and that confidence does not properly reflect predictive accuracy (Oskamp, 1965). This may help explain why, when we have obtained a rich and consistent impression of a person, we often feel that we "understand" the person and can foresee his future behavior. Because this feeling often persists even when we know that our predictions are generally inaccurate, it may be termed the <u>illusion of validity</u>. When interviewing a candidate for a position, for example, many of us find it impossible to resist this illusion, despite our knowledge that interviews are notoriously fallible.

In summary, intuitive predictions appear insensitive to considerations of predictive accuracy which moderate predictions in the normative theory. Even when expected accuracy is low, people may predict infrequent events and extreme values with high confidence, if these outcomes are representative of their fallible impressions. These observations of widespread and tenacious biases raise the intriguing questions of why people do not extract the correct intuitions from lifelong experience, and of how, if at all, the correct intuitions could be acquired.

References

Beyth, R. Man as a weak intuitive statistician: On erroneous intuitions concerning the description and prediction of events. Unpublished Master's Thesis (in Hebrew), The Hebrew University, Jerusalem, 1972.

Campbell, D. T. Reforms as experiments. American Psychologist, 1969, 24, 409-429.

Chapman, L. J., & Chapman, J. P. Genesis of popular but erroneous psychodiagnostic observations. Journal of Abnormal Psychology, 1967, 73, 193-204.

Gray, C. W. Predicting with intuitive correlations. Psychonomic Science, 1968, 11, 41-43.

Johnson, D. M. Systematic Introduction to the Psychology of Thinking. New York: Harper & Row, 1972.

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

Mischel, W. Personality and Assessment. New York: Wiley, 1968.

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

Slovic, P. Cue consistency and cue utilization in judgment. American Journal of Psychology, 1966, 79, 427-434.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Tversky, A., & Kahneman, D. Belief in the law of small numbers. Psychological Bulletin, 1971, 76, 105-110.

Wallis, W. A., & Roberts, H. V. Statistics: A New Approach. New York: The Free Press, 1956.

Footnotes

1. Research for this study was supported by the following grants: Grants MH 12972 and MH 21216 from the National Institute of Mental Health and Grant RR 05612 from the National Institutes of Health, U.S. Public Health Service; Grant GS 3250 from the National Science Foundation; Computing assistance was obtained from the Health Sciences Computing Facility, UCLA, sponsored by Grant MH 10822 from the U.S. Public Health Service.

We thank Robyn Dawes, Lewis Goldberg, and Paul Slovic for their comments. Sundra Gregory and Richard Kleinknecht assisted in the preparation of the test material and the collection of data.

2. Unless otherwise specified, the subjects in the studies reported in this paper were paid volunteers recruited through a student paper at the University of Oregon. Data were collected in group settings.

3. In computing this correlation, the ranks were inverted so that a high judged likelihood was assigned a high value.

4. For enlightening discussions of regression fallacies in research see e.g., Campbell (1969), Wallis and Roberts (1956).

v. 12, no. 5 was never published.