Judgment under uncertainty: Heuristics and biases

Edited by

Daniel Kahneman
University of British Columbia

Paul Slovic
Decision Research
A Branch of Perceptronics, Inc.
Eugene, Oregon

Amos Tversky
Stanford University

Cambridge University Press

Cambridge
London New York New Rochelle
Melbourne Sydney

4. On the psychology of prediction

Daniel Kahneman and Amos Tversky

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. Two classes of prediction are discussed: category prediction and numerical prediction. In a categorical case, the prediction is given in nominal form, for example, the winner in an election, the diagnosis of a patient, or a person's future occupation. In a numerical case, the prediction is given in numerical form, for example, the future value of a particular stock or of a student's grade point average.

In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors (Kahneman & Tversky, 1972b, **3**; Tversky & Kahneman, 1971, **2**; 1973, **11**). The present paper is concerned with the role of one of these heuristics – representativeness – in intuitive predictions.

Given specific evidence (e.g., a personality sketch), the outcomes under consideration (e.g., occupations or levels of achievement) can be ordered by the degree to which they are representative of that evidence. The thesis of this paper is that people predict by representativeness, that is, they select or order outcomes by the degree to which the outcomes represent the essential features of the evidence. In many situations, representative outcomes are indeed more likely than others. However, this is not always the case, because there are factors (e.g., the prior probabilities of outcomes and the reliability of the evidence) which affect the likelihood of outcomes but not their representativeness. Because these factors are ignored, intuitive predictions violate the statistical rules of prediction in systematic and fundamental ways. To confirm this hypothesis, we show that the ordering of outcomes by perceived likelihood coincides with their ordering by representativeness and that intuitive predictions are essentially unaffected by considerations of prior probability and expected predictive accuracy.

In the first section, we investigate category predictions and show that they conform to an independent assessment of representativeness and that they are essentially independent of the prior probabilities 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 following three sections discuss, in turn, methodological issues in the study of prediction, the sources of unjustified confidence in predictions, and some fallacious intuitions concerning regression effects.

This chapter originally appeared in *Psychological Review*, 1973, *80*, 237–251. Copyright © 1973 by the American Psychological Association. Reprinted by permission.

Categorical prediction

Base rate, similarity, and likelihood

The following experimental example illustrates prediction by representativeness and the fallacies associated with this mode of intuitive prediction. A group of 69 subjects[1] (the *base-rate* group) was asked the following question: "Consider all first-year graduate students in the U.S. today. Please write down your best guesses about the percentage of these students who are now enrolled in each of the following nine fields of specialization." The nine fields are listed in Table 1. The first column of this table presents the mean estimates of base rate for the various fields.

A second group of 65 subjects (the *similarity* group) was presented with the following personality sketch:

> 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 subjects were asked to rank the nine areas in terms of "how similar is Tom W. to the typical graduate student in each of the following nine fields of graduate specialization?" The second column in Table 1 presents the mean similarity ranks assigned to the various fields.

Finally, a *prediction* group, consisting of 114 graduate students in

[1] 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.

Table 1. *Estimated base rates of the nine areas of graduate specialization and summary of similarity and prediction data for Tom W.*

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

psychology at three major universities in the United States, was given the personality sketch of Tom W., with the following additional information:

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 each of these fields.

The third column in Table 1 presents the means of the ranks assigned to the outcomes by the subjects in the prediction group.

The product-moment correlations between the columns of Table 1 were computed. The correlation between judged likelihood and similarity is .97, while the correlation between judged likelihood and estimated base rate[2] is −.65. Evidently, judgments of likelihood essentially coincide with judgments of similarity and are quite unlike the estimates of base rates. This result provides a direct confirmation of the hypothesis that people predict by representativeness, or similarity.

The judgments of likelihood by the psychology graduate students drastically violate the normative rules of prediction. More than 95% of those 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

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

humanities or education against computer science are about 3 to 1. (The actual odds are considerably higher.)

According to Bayes' rule, it is possible to overcome the prior odds against Tom W. being in computer science rather than in humanities or education, if the description of his personality is both accurate and diagnostic. The graduate students in our study, however, 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 on a description derived from such tests and ignored the base rates.

In general, three types of information are relevant to statistical prediction: (a) prior or background information (e.g., base rates of fields of graduate specialization); (b) specific evidence concerning the individual case (e.g., the description of Tom W.); (c) the expected accuracy of prediction (e.g., the estimated probability of hits). A fundamental rule of statistical prediction is that expected accuracy controls the relative weights assigned to specific evidence and to prior information. When expected accuracy decreases, predictions should become more regressive, that is, closer to the expectations based on prior information. In the case of Tom W., expected accuracy was low, and prior probabilities should have been weighted heavily. Instead, our subjects predicted by representativeness, that is, they ordered outcomes by their similarity to the specific evidence, with no regard for prior probabilities.

In their exclusive reliance on the personality sketch, the subjects in the prediction group apparently ignored the following considerations. First, given the notorious invalidity of projective personality tests, it is very likely that Tom W. was never in fact as compulsive and as aloof as his description suggests. Second, even if the description was valid when Tom W. was in high school, it may no longer be valid now that he is in graduate school. Finally, even if the description is still valid, there are probably more people who fit that description among students of humanities and education than among students of computer science, simply because there are so many more students in the former than in the latter field.

Manipulation of expected accuracy

An additional study tests the hypothesis that, contrary to the statistical model, a manipulation of expected accuracy does not affect 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 in the context of a longitudinal study. 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 judgments, 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 cases had been randomly selected from among the participants in the original study who are now first-year graduate students. 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 group ($t = 3.72$, $p < .001$). However, the orderings of the nine outcomes produced under the low-accuracy instructions were not significantly 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 averages of these individual correlations were .13 for subjects in the high-accuracy group and .16 for subjects in the low-accuracy group. The difference does not approach significance ($t = .42$, $df = 103$). 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 expected accuracy had no effect on predictions, the two prediction groups were pooled. Subsequent analyses were the same as in the Tom W. study. For each description, two correlations were computed: (a) between mean likelihood rank and mean similarity rank and (b) between mean likelihood rank 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 high. In contrast, there is 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.

Table 2. *Product-moment correlations of mean likelihood rank with mean similarity rank and with base rate*

	Modal first prediction				
	Law	Computer science	Medicine	Library science	Business administration
With mean similarity rank	.93	.96	.92	.88	.88
With base rate	.33	.35	.27	−.03	.62

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-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.

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

Prior versus individuating evidence

The next study provides a more stringent test of the hypothesis that intuitive predictions are dominated by representativeness and are relatively insensitive to prior probabilities. In this study, 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). Subjects in another group (the high-engineer, H group; $N = 86$) were given identical instructions except for the prior probabilities: they were told that the set 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 both the high-engineer and low-engineer groups, half of the subjects were asked to evaluate, for each description, the probability that the person described was an engineer (as in the example above), while the other subjects evaluated, for each description, the probability that the person described was 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 each description. Consequently, the data for the two forms were pooled, and the results are presented 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, $O = Q \cdot R$, where Q denotes the prior odds that a randomly selected description belongs to an engineer rather than to a lawyer; and R is the likelihood ratio for a particular description, that is, 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 a population of lawyers will be so described.

For the high-engineer group, who were told that the sample consists of 70 engineers and 30 lawyers, the prior odds Q_H equal 70/30. For the low-engineer group, the prior odds Q_L equal 30/70. Thus, for each description, the ratio of the posterior odds for the two groups is

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

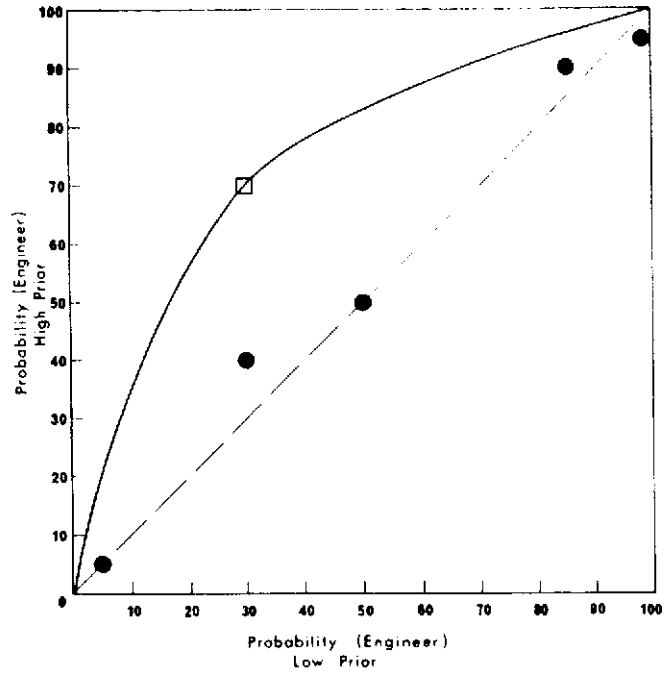

Figure 1. Median judged probability (engineer) for five descriptions and for the null description (square symbol) under high and low prior probabilities. (The curved line displays the correct relation according to Bayes's rule.)

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.

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 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 the low-engineer group and 55% for the high-engineer group ($t = 3.23$, $df = 169$, $p < .01$). 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 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 evidence. 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 low- and high-engineer groups (see Figure 1). The contrast between the responses to this description and to the null description is illuminating. Evidently, people respond differently when given no specific evidence and when given worthless evidence. When no specific evidence is given, the prior probabilities are properly utilized; when worthless specific evidence is given, prior probabilities are ignored.[3]

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 nonrecipients of fellowships. Rather, it is natural to regard the outcome (i.e., obtaining a fellowship) as determined by a cutoff point along the dimension of academic achievement or ability. Prior probabilities, that is, the percentage of fellowships in the relevant group, could be used to define the outcomes by locating the cutoff point. Consequently, they are not likely to be ignored. In addition, we would expect extreme prior probabilities to have some effect even in the presence of clear stereotypes of the outcomes. A precise delineation of the conditions under which prior information is used or discarded awaits further investigation.

One of the basic principles of statistical prediction is that prior probability, which summarizes what we knew about the problem before receiving independent 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,

[3] But see p. 159.

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.

Numerical prediction

A fundamental rule of the normative theory of prediction is that the variability of predictions, over a set of cases, should reflect predictive accuracy. When predictive accuracy is perfect, one predicts the criterion value that will actually occur. When uncertainty is maximal, a fixed value is predicted in all cases. (In category prediction, one predicts the most frequent category. In numerical prediction, one predicts the mean, the mode, the median, or some other value depending on the loss function.) Thus, the variability of predictions is equal to the variability of the criterion when predictive accuracy is perfect, and the variability of predictions is zero when predictive accuracy is zero. With intermediate predictive accuracy, the variability of predictions takes an intermediate value, that is, predictions are regressive with respect to the criterion. Thus, the greater the uncertainty, the smaller the variability of predictions. Predictions by representativeness do not follow this rule. It was shown in the previous section that people did not regress toward more frequent categories when expected accuracy of predictions was reduced. The present section demonstrates an analogous failure in the context of numerical prediction.

Prediction of outcomes versus evaluation of inputs

Suppose one is told that a college freshman has been described by a counselor as intelligent, self-confident, well-read, hard working, and inquisitive. Consider two types of questions that might be asked about this description:

(*a*) *Evaluation:* How does this description impress you with respect to academic ability? What percentage of descriptions of freshmen do you believe would impress you more? (*b*) *Prediction:* What is your estimate of the grade point average that this student will obtain? What is the percentage of freshmen who obtain a higher grade point average?

There is an important difference between the two questions. In the first, you evaluate the input; in the second, you predict an outcome. Since there is surely greater uncertainty about the second question than about the first, your prediction should be more regressive than your evaluation. That is, the percentage you give as a prediction should be closer to 50% than the percentage you give as an evaluation. To highlight the difference between the two questions, consider the possibility that the description is

inaccurate. This should have no effect on your evaluation: the ordering of descriptions with respect to the impressions they make on you is independent of their accuracy. In predicting, on the other hand, you should be regressive to the extent that you suspect the description to be inaccurate or your prediction to be invalid.

The representativeness hypothesis, however, entails that prediction and evaluation should coincide. In evaluating a given description, people select a score which, presumably, is most representative of the description. If people predict by representativeness, they will also select the most representative score as their prediction. Consequently, the evaluation and the prediction will be essentially identical. 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, allegedly 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. 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 grade point average 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 2, which plots, for each description, the mean prediction of percentile grade point average against the mean evaluation. The only systematic discrepancy between predictions and evaluations is observed in the adjectives 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 ($N = 36$) ($t = 1.25$, $df = 72$, *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 ($N = 63$) ($t = .75$, $df = 98$, *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

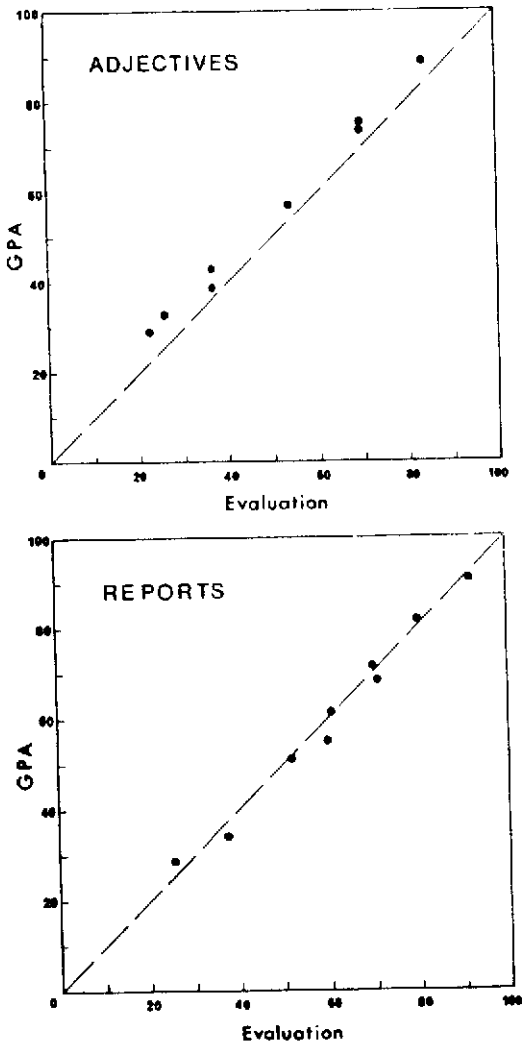

Figure 2. Predicted percentile grade point average as a function of percentile evaluation for adjectives and reports.

merely evaluated the impression obtained from each description. In the statistical theory of prediction, the observed equivalence between prediction and evaluation would be justified only if predictive accuracy were perfect, a condition which could not conceivably be met in these studies.

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 the quality of the lesson described in the paragraph, in percentile scores relative to the stated population. Subjects in a prediction group were asked to predict in percentile scores the current standing of each teacher, that is, his overall 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-subjects 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.

Prediction versus translation

The previous studies showed that predictions of a variable are not regressive when compared to evaluations of the inputs in terms of that variable. In the following study, we show that there are situations in which predictions of a variable (academic achievement) are no more regressive than a mere translation of that variable from one scale to another. The grade point average was chosen as the outcome variable, because it correlates and distributional properties are well known to the subject population.

Three groups of subjects participated in the experiment. Subjects in all groups predicted the grade point average of 10 hypothetical students on the basic 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 grade point average.* 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 grade point average for that year." It was explained to the subjects that "a percentile score of 65, for example, means that the grade point average 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."

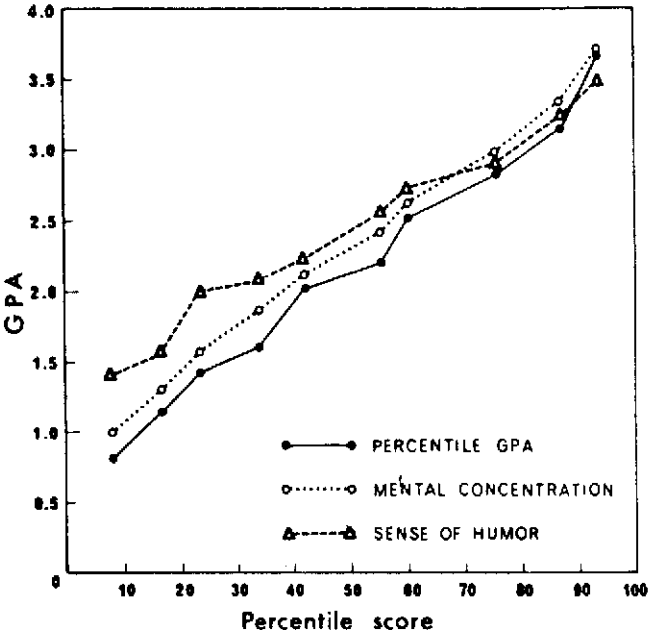

Figure 3. Predictions of grade point average from percentile scores on three variables.

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 grade point average than students who score low. However, it is not possible to predict grade point average from sense of humor with high accuracy."

In the present design, all subjects predicted grade point average on the basis of the same set of percentile scores. Group 1 merely translated values of percentile grade point average onto the grade point average scale. Groups 2 and 3, on the other hand, predicted grade point average from more remote inputs. Normative considerations therefore dictate that the predictions of these groups should be more regressive, that is, less variable, than the judgments of Group 1. The representativeness hypothesis, however, suggests a different pattern of results.

Group 2 predicted from a potentially valid but unreliable test of mental concentration which was presented as a measure of academic ability. We hypothesized that the predictions of this group would be nonregressive when compared to the predictions of Group 1. In general, we conjecture that the achievement score (e.g., grade point average) which best repre-

Table 3. *Averages of individual prediction statistics for the three groups and results of planned comparisons between groups 1 and 2, and between groups 2 and 3*

Statistic	Group 1. Percentile grade point average	1 vs. 2	2. Mental concentration	2 vs. 3	3. Sense of humor
Mean predicted grade point average	2.27	*ns*	2.35	.05	2.46
SD of predictions	.91	*ns*	.87	.01	.69
Slope of regression	.030	*ns*	.029	.01	.022
r	.97	*ns*	.95	*ns*	.94

sents a percentile value on a measure of ability (e.g., mental concentration) is that which corresponds to the same precentile on the scale of achievement. Since representativeness is not affected by unreliability, we expected the predictions of grade point average from the unreliable test of mental concentration to be essentially identical to the predictions of grade point average from percentile grade point average. The predictions of Group 3, on the other hand, were expected to be regressive because sense of humor is not commonly viewed as a measure of academic ability.

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

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 grade point average on the input scores, and the product–moment correlation between them. The means of these values for the three groups are shown in Table 3.

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 3. There are no significant differences between the predictions from percentile grade point average and from mental concentration. Thus, people fail to regress when predicting a measure of achievement from a measure of ability, however unreliable.

The predictions from sense of humor, on the other hand, are regressive, although not enough. The correlation between grade point average and sense of humor inferred from a comparison of the regression lines is about .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 grade point average. We have observed this finding in many studies. When predicting the academic achievement of an individual on the basis of imperfect information, subjects exhibit leniency (Guilford, 1954). They respond to a reduction of validity by raising the predicted level of performance.

Predictions are expected to be essentially nonregressive whenever the input and outcome variables are viewed as manifestations of the same trait. An example of such predictions has been observed in a real-life setting, the Officer Selection Board of the Israeli Army. The highly experienced officers who participate in the assessment team normally evaluate candidates on a 7-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 200 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 keenly aware of research indicating that their predictive validity was only moderate (on the order of .20 to .40). Nevertheless, their predictions were nonregressive.

Methodological considerations

The representativeness hypothesis states that predictions do not differ from evaluations or assessments of similarity, although the normative statistical theory entails that predictions should be less extreme than these judgments. The test of the representativeness hypothesis therefore requires a design in which predictions are compared to another type of judgment. Variants of two comparative designs were used in the studies reported in this paper.

In one design, labeled *A-XY*, different groups of subjects judged two variables (*X* and *Y*) on the basis of the same input information (*A*). In the case of Tom W., for example, two different groups were given the same input information (*A*), that is, a personality description. One group ranked the outcomes in terms of similarity (*X*), while the other ranked the outcomes in terms of likelihood (*Y*). Similarly, in several studies of numerical prediction, different groups were given the same information (*A*), for example, a list of five adjectives describing a student. One group provided an evaluation (*X*) and the other a prediction (*Y*).

In another design, labeled *AB-X*, two different groups of subjects judged the same outcome variable (*X*) on the basis of different information inputs

(*A* and *B*). In the engineer/lawyer study, for example, two different groups made the same judgment (*X*) of the likelihood that a particular individual is an engineer. They were given a brief description of his personality and different information (*A* and *B*) concerning the base-rate frequencies of engineers and lawyers. In the context of numerical prediction, different groups predicted grade point average (*X*) from scores on different variables, percentile grade point average (*A*) and mental concentration (*B*).

The representativeness hypothesis was supported in these comparative designs by showing that contrary to the normative model, predictions are no more regressive than evaluations or judgments of similarity. It is also possible to ask whether intuitive predictions are regressive when compared to the actual outcomes, or to the inputs when the inputs and outcomes are measured on the same scale. Even when predictions are no more regressive than translations, we expect them to be slightly regressive when compared to the outcomes, because of the well-known central-tendency error (Johnson, 1972; Woodworth, 1938). In a wide variety of judgment tasks, including the mere translation of inputs from one scale to another, subjects tend to avoid extreme responses and to constrict the variability of their judgments (Stevens & Greenbaum, 1966). Because of this response bias, judgments will be regressive, when compared to inputs or to outcomes. The designs employed in the present paper neutralize this effect by comparing two judgments, both of which are subject to the same bias.

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 previous studies of prediction have dealt with the learning of functional or statistical relations among variables with which the subjects had no prior acquaintance. These studies typically involve a large number of trials and various forms of outcome feedback. (Some of this literature has been reviewed in Slovic and Lichtenstein, 1971.) In studies of repetitive predictions with feedback, subjects generally predict by selecting outcomes so that the entire sequence or pattern of predictions is highly representative of the distribution of 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, that is, the joint distribution of inputs and outcomes (see, e.g., Gray, 1968). To do so, subjects resort to a mixed strategy: for any given input they generate a distribution of different predictions. These predictions reflect 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 representing 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 rule: 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.

Confidence and the illusion of validity

As demonstrated in the preceding sections, one predicts by selecting the outcome that is most representative of the input. We propose that the degree of confidence one has in a prediction reflects the degree to which the selected outcome is more representative of the input 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 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 B grades in two separate introductory courses than on the basis of an A and a C. Indeed, internal variability 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 model of prediction (i.e., 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.

To demonstrate this effect, we required subjects to predict grade point average on the basis of two pairs of aptitude tests. Subjects were told that one pair of tests (creative thinking and symbolic ability) was highly correlated, while the other pair of tests (mental flexibility and systematic reasoning) was not correlated. The scores they encountered conformed to these expectations. (For half of the subjects the labels of the correlated and the uncorrelated pairs of tests were reversed.) Subjects were told that "all tests were found equally successful in predicting college performance." In

this situation, of course, a higher predictive accuracy can be achieved with the uncorrelated than with the correlated pair of tests. As expected, however, subjects were more confident in predicting from the correlated tests, over the entire range of predicted scores ($t = 4.80$, $df = 129$, $p < .001$). 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 Johnson, 1972). Subjects predict outstandingly high achievement with very high confidence, and they have more confidence in the prediction of utter failure than of 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 analysis shows that the factors which enhance confidence, for example, consistency and extremity, are often negatively correlated with predictive accuracy. Thus, people are prone to experience much confidence in highly fallible judgments, a phenomenon that may be termed the *illusion of validity*. Like other perceptual and judgmental errors, the illusion of validity often persists even when its illusory character is recognized. When interviewing a candidate, for example, many of us have experienced great confidence in our prediction of his future performance, despite our knowledge that interviews are notoriously fallible.

Intuitions about 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 a major source of difficulty is that regression effects typically violate the intuition that the predicted outcome should be maximally representative of the input information.[4]

To illustrate the persistence of nonregressive 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 than 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 counterregressive 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.[5] 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 a 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 variances, the following two statements are logically equivalent: (a) Y is regressive with respect to X; (b) 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

[4] The expectation that every significant particle of behavior is highly representative of the actor's personality may explain why laymen and psychologists alike are perennially surprised by the negligible correlations among seemingly interchangeable measures of honesty, of risk taking, of aggression, and of dependency (Mischel, 1968).

[5] For enlightening discussions of regression fallacies in research, see, for example, Campbell (1969) and Wallis and Roberts (1956).

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 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 suggested that regression could cause 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 discussed 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.

5. Studies of representativeness

Maya Bar-Hillel

Daniel Kahneman and Amos Tversky have proposed that when judging the probability of some uncertain event people often resort to heuristics, or rules of thumb, which are less than perfectly correlated (if, indeed, at all) with the variables that actually determine the event's probability. One such heuristic is *representativeness*, defined as a subjective judgment of the extent to which the event in question "is similar in essential properties to its parent population" or "reflects the salient features of the process by which it is generated" (Kahneman & Tversky, 1972b, p. 431, 3). Although in some cases more probable events also appear more representative, and vice versa, reliance on the representativeness of an event as an indicator of its probability may introduce two kinds of systematic error into the judgment. First, it may give undue influence to variables that affect the representativeness of an event but not its probability. Second, it may reduce the importance of variables that are crucial to determining the event's probability but are unrelated to the event's representativeness.

The representativeness concept has occasionally been criticized as too vague and elusive, presumably because it lacks a general operational definition. This is not to say, however, that it is impossible to assess representativeness independently of probability judgments, a conclusion which has often been implied by the critics. In the "Tom W." study, for example, Kahneman and Tversky (1973, 4) defined representativeness as the similarity of some individual, Tom W., to "the typical graduate student in . . . [some] fields of graduate specialization" (1973, p. 238) and ranked it independently of the likelihood that Tom W. was enrolled in those fields. In other studies, the independent ranking by representativeness was sidestepped only because readers could so readily supply it themselves via thought experiments.