The Prediction of Social and Technological Events

BY A. KAPLAN, A. L. SKOGSTAD, AND M. A. GIRSHICK

A group of individuals with a generally high education level was asked to make predictions concerning a large number of future events in order to investigate certain aspects of the use of expert opinion in policy making. In particular, the following questions were explored: how good are expert predictions in areas germane to policy; how can such predictions be approved; and how can the reliability of a given opinion be appraised beforehand? While the design of the study makes projection of its results on other situations questionable, it throws considerable light on the problems involved. Among other results, it was found that confidence in prediction does not necessarily show a correlation with success in prediction, that predictions made by groups of people are more likely to be right than predictions made by the same individuals working alone, and that the reliability of predictions can be appraised to some extent by examining the character of the justifications given for them.

Dr. Kaplan is Associate Professor of Philosophy at U.C.L.A.; Dr. Girshick is Professor of Mathematics at Stanford; and Mrs. Skogstad is an economist with The RAND Corporation, Santa Monica, California.

Policy making rests in part on anticipation of the future—predictions of the conditions which policy must face, and of the consequences of and responses to alternative lines of action.[1]

Many policy decisions require foreknowledge of events which cannot be forecast either by strict causal chains (as can eclipses) or by stable statistical regularities (as can the number of traffic deaths in a given period). For prediction of such events, the policy maker has no recourse but reliance on the judgment of experts. In particular, forecasts of many political and economic changes and scientific and technological developments often fall in this category.

The policy maker, though he may not himself be an expert on the particular field basic to the decision, must nevertheless evaluate the expert's judgment, appraise its likelihood of success, and attempt to improve on it. Such tasks are involved in the selection of experts, in the choice between conflicting opinions of experts, and in the decision of how much reliance to put on the selected opinion. Experts can and do perform these operations on each other's predictions, but an ultimate judgment on these matters remains an inescapable part of the policy maker's function.

[1] In part, of course, recommendations of policy also depend on such factors as the cost and feasibility of one course of action as against another, or the degree of conformity of the available alternatives to the present values or past decisions of the policy maker.

DEFINING THE PROBLEM

This paper reports on an exploratory study of certain aspects of the following three questions, among the many that can be raised concerning predictions and their use in policy formation:[2]

1. The problem of evaluation: How good are expert predictions in areas germane to policy?
2. The problem of improvement: How can such predictions be improved?
3. The problem of appraisal: How can the reliability of a given prediction be appraised beforehand?

Moreover, we are interested here in only a few of the many factors of the problem of prediction. These factors are, of course, difficult to separate; and in any fully adequate treatment of the problem they must be considered in their interaction.

The Problem of Evaluation

Taken in the concrete, the evaluation of predictions of the expert adviser to the policy maker requires consideration of the entire complex making up the politics of the advisory function. The expert is here playing a role in the political process, and this fact has a great bearing on the predictions made. To what extent does political tact result in suppression of predictions critical of other participants in the political arena? How far does political prudence shape the prediction in terms of expected effects of the prediction's being made at all? How widely is the "principle of the oracle" (non-falsifiable prediction) followed, in order to secure the expert's influence against the adverse effects of repeated failures? The task of evaluation also presents us with a variety of psychological problems: the effect of emotional involvement in what is being predicted; the role of knowledge, as theory and as information about matters of fact; the part played by anxiety level, optimism-pessimism, and other characteristics of the personality.

Most of these factors, important as they are, are not dealt with to any extent in this paper. From the total complex situation we abstract certain relatively simple components for analysis. How successful *are* predictions of social and scientific change? Do these two subject matters differ stably in predictability? How precisely can the probability of events in these areas be specified? What differences can be observed in the predictability of events in the relatively near and more distant futures?

The Problem of Improvement

Here again we consider only one or two of the many factors involved in the concrete situation. The problem of improvement of prediction obviously raises questions as to optimal recruitment of advisers. The notion of "expert" is primarily, in its application, a sociological rather than logical category, depending as much as it does on reputation, influence, skills in managing interpersonal relations, etc.

Apart from suitable selection of predictors, improvement of prediction rests also on developments in the underlying theory and in techniques of observing and refining data. And further improvements might be sought through analy-

[2] The authors wish to thank Drs. A. Mood, O. Helmer, and others who gave indispensable assistance in carrying out this study.

sis of the conditions implicit in a given prediction, by differentiating the diagnosis made from the prognosis resting upon it, and by correcting the one or the other as the evidence indicates.

This paper, however, deals with quite another dimension of the problem of improvement. By how much can the reliability of prediction be enhanced by taking statistical combinations of the forecasts of individual predictors—the forecast of the majority of a panel, for example, or the mean of their estimates of the probability of specified alternatives? Can prediction be improved by a "consensus" statistic which assigns weights to the predictions of the individual experts in terms of past performance? Can predictions be improved if they are made collectively by a group rather than by a number of predictors working independently?

The Problem of Appraisal

Whether or not prediction can be improved, a great deal would be gained if it were at any rate possible to appraise correctly the likelihood that a given prediction will be fulfilled. The problem is essentially that of specifying sub-populations of predictions in which the probability of success remains relatively stable. Such a specification plainly involves consideration of the psychological and political variables already mentioned. A firm prediction made by a predictor who is habitually timid presumably has a different likelihood of success than if made by one who is habitually overconfident. Private predictions made in the process of policy formation are to be appraised differently than those made for public or foreign consumption. And so on.

We confine ourselves here to the possibilities of appraising predictions in terms of supplementary information made available by the predictor. Are predictions made with high confidence more likely on the whole to be successful than those made with lesser confidence? What is the effect of the rank order of likelihood assigned to the predicted alternative in a given set of alternatives? What appraisal can be made in terms of the specific probability assigned the alternative? How adequately can a prediction be appraised in terms of the basis on which it rests, as stated by the predictor?

These, then, are the questions with which this paper is concerned. They have been placed in the wider setting of the whole problem, not only to make clear the limitations of the study here reported on, but also in the hope of stimulating interest in other aspects of the problem. For, in the opinion of the writers, the topic of prediction has received far too little attention in the light of its importance, both practical and theoretical. With the exception of a few papers, available materials are limited to the study of opinion.[3] For the most part, it is public opinion which is in question, but even expert opinion is usually dealt with as an expression of a point of view rather than as verifiable prediction. Indeed, the polls rarely distinguish between verifiable predictions about matters of fact and unverifiable judgments of value. Prediction is taken into account only when the opinion

[3] Among the few studies of prediction concerned with the *predictive* validity of opinion are D. McGregor's "The Major Determinants of the Prediction of Social Events" and Hadley Cantril's "The Prediction of Social Events," both in the *Journal of Abnormal Psychology,* 33 (1938).

studied is one that is thought to warrant inferences about special events (e.g., elections); and in that case, the prediction studied is the poll-taker's own, not that of the persons polled. Though much is known about the prediction of public opinion itself, and perhaps something about the prediction of social events from a knowledge of opinion, the general problem of prediction as presented above is largely unexplored.

DESIGN OF THE EXPERIMENT

This paper reports on a pilot study concerning the three basic problems of evaluation, improvement, and appraisal. A group of 26 predictors was used. This group included the following persons: fifteen mathematicians and statisticians, four engineers, four individuals trained in economics or business administration, one office manager, one secretary, and one writer. Only two did not have a college education; thirteen spent six years or more in college or graduate school; eight of these had a Ph.D. degree. In addition, the group averaged over seven years of experience (teaching, research, or practice) in the field of specialization. Direct measures of intelligence and personality characteristics were not available; "general knowledge" in the fields of prediction was, however, measured, as will be described below.

Predictions were made by filling out questionnaires. Each consisted of ten to sixteen questions, evenly divided in subject-matter between "social science" and "natural science" events. The former called for predictions on domestic politics, foreign politics, and economic affairs; the latter called for predictions on the physical sciences, the life sciences, and technology. The predictions were all categorical rather than conditional—i.e., of the form "The likelihood that E will occur is p," rather than "If the conditions C are satisfied the likelihood of E is p." The predictions also concerned future events on which the predictors could have no effect or a negligible effect—i.e., the predictions were not of responses to their own acts. Sample questions are given in Appendix A.

In all, 152 questions were asked in 13 questionnaires. Of these, 29 were rejected before the analysis for the following reasons:

Verification impossible	10
Incorrect information given	7
Ambiguous wording	6
Alternatives not exhaustive	4
Verified before questionnaire issued	2
TOTAL	29

The responses to a total of 123 questions, comprising 3007 separate predictions, were thus available for analysis.

In each question the predictor was offered four exhaustive and exclusive alternative outcomes with a time limit for occurrence of the predicted event set at 20 weeks (or less) from the date of the questionnaire.[4] The predictor

[4] The attempt was made to select alternatives that represented "real possibilities"—i.e., with antecedent probabilities as high as possible. For instance, on the question of whom the Republicans would nominate for President, the alternatives given were Dewey, Stassen, Taft, and "Other," simply because these were judged by the experimenters to be genuine alternatives. But, of course, there was no objective check on such judgments. The selection of four "real possibilities" was often very difficult. We cannot say what effect this had on the results.

was required to give for each alternative his judgment of "the likelihood of its occurrence," expressed as a value from 0 to 100, inclusive. He was also told that the values for the four alternatives of each question were to add to 100.

In addition, space was provided for recording "Basis for Your Judgment." No instructions were given as to how this space was to be filled.

A new questionnaire was distributed weekly for 13 weeks to all the predictors simultaneously. They were given three hours in which to answer the questionnaire without any research, and, with the exceptions described in the next paragraph, without discussion with other predictors.

Thirteen of the 26 predictors constituted a special set for the study of group predictions. They were divided each week into three quartets (provision being made for one absentee). One quartet, known as the *independent group*, answered the questionnaire individually, as usual; this was the control group. One quartet, the *cooperative group*, discussed the questions together, then answered them individually. The third, the joint group, discussed the questions, then came to some collective decision, giving one answer for the entire group. This phase was so designed that every individual participated four times in each of the three types of quartets, working together once with each of the other twelve predictors in this phase.[5]

Before stating the results of this study, we wish to emphasize an important point in connection with their interpretation. *There is no assurance that these results concern a statistically stable population.* However, it is encouraging that when the analysis was carried out on the first 72 questions to be verified, then repeated on the full set of 123, the results were substantially the same. But it would be hazardous to infer that our basic conclusions would be unaltered by a study with different types of questions, a questionnaire of different design, and especially with different predictors. The pilot study here reported was primarily directed at solving problems of method. Its authors believe the results to be significant only in so far as they are suggestive of hypotheses to guide further research. By themselves they are not believed to provide an adequate basis for extrapolation to other predictions.

A. THE PROBLEM OF EVALUATION

1. Success

We call a prediction *successful* when the alternative to which the highest value was assigned is in fact verified. The success attained by the predictors in this study ranged from 71 per cent to 28 per cent, around a mean of 53 per cent. Seven of the 26 predictors scored above 60 per cent; three, below 40 per cent. Of course, since four alternatives were presented in each question, a mechanically random process would have scored a mean success of 25 per cent. Since, however, the four alternatives cannot by any means be regarded as having equal antecedent probabilities, this standard of comparison is misleading. Plainly, a success as high as desired could have been artificially obtained by formulating alternatives with extremely high or low antecedent prob-

[5] The authors wish to thank Dr. J. Youden for the design of this phase of the experiment.

abilities. In formulating the questions for this study, the attempt was made to select the alternatives actually posed in the daily and scientific press, and to choose questions of the sort that might actually confront policy makers. How close we came to this objective can be judged only impressionistically by an inspection of the questions (see Appendix A).

The evaluation of a predictor's success must also take into account the *definiteness* of his prediction, i.e., the degree to which he committed himself on the relative likelihood of the four alternatives. We may measure definiteness by the extent to which the values assigned per question differed from 25 for each alternative. If the values were equally distributed, the prediction would be perfectly indefinite; the maximal definiteness consists in assigning one alternative 100 and zero to the other three. Accordingly, the square root of the mean square deviation from 25 for the values assigned in each question was computed for each predictor. The result ranged from 12 to 35, with a mean of 27, the maximum possible range being from 0 to 43.

What is noteworthy is that definiteness in the sense we are using it does not show a significant correlation with success (rank correlation $r = .2$). Predictors who were often right were, on the whole, scarcely more definite in their predictions than those who were often in error. If a predictor's definiteness could be taken to express his own appraisal of his prediction, this would suggest that ability to predict and ability to appraise one's own predictions are distinct components of expertness. However, the measure of definiteness used here may be reflecting a fairly even distribution of the antecedent probabilities of the four alternatives. A predictor might be indefinite, that is, not because he cannot appraise his prediction, but because he is correctly estimating the four alternatives to have approximately equal probabilities. This is the question to which we next turn.

2. Precision

If the values assigned to the various alternatives were intended by the predictors to express probabilities, then to assign a value x to an alternative would be to judge that it belongs to a class of predictions of which x per cent will be verified.[6] We say that a predictor is *precise* in the degree to which his assignment of values satisfies this condition—so that, for instance, if he were perfectly precise, of all the alternatives to which he assigned a value of, say, 60, in fact 60 per cent would be verified. The following definitions of "precision" were used:

$$\rho_1 = \sqrt{\Sigma_i\ (S_i/V_i - 1)^2\ F_i} \qquad (i = 25, 26, \ldots, 100)$$

$$\rho_2 = \sqrt{\Sigma_i\ (S_i - V_i)^2\ F_i} \qquad (i = 0, 1, \ldots, 100)$$

where S_i is the per cent of all predictions made with the value V_i which were verified, and F_i is the relative frequency with which V_i was used. ρ_1 was applied to the top-ranking values among each set of alternatives, ρ_2 to all values.

When the precision of a predictor, as here defined, is correlated with his success, the results are .6 and .5, re-

[6] It is an open question whether such was their intention, since the instructions to the predictors deliberately did not specify any interpretation of "likelihood of occurrence."

TABLE 1

PREDICTION BY TIME INTERVALS

*Time Interval in Days**	*Number of Questions*	*Number of Predictions*	*Success (Per Cent)*	*Frequency of Values Above 50 (Per Cent)*	*Success in Values Above 50 (Per Cent)*
0 - 48	14	301	49%	50%	58%
49 - 93	16	348	45	51	57
94 - 139	9	191	35	50	43
140	84	1813	55	54	69

* Number of days between date of questionnaire and date of verifying event.

spectively. That is, predictors whose preferred alternatives were often verified were rather likely to assign their values more "precisely" than the less successful predictors. Their values might be said to provide somewhat more accurate estimates of the probabilities of the events in question.

Extreme errors in precision might be called *blunders*—assignments of very high values to alternatives that do not occur, or very low ones to those that do.[7] (Because every case of the first type of error is also one of the second, since the values of the four alternatives in each question must add to 100, we report only on the second type of error.) The frequency of such "blunders" by each predictor ranged from 8 per cent to 48 per cent of the questions, with a mean of 20 per cent for the whole set of predictors. Infrequency of blunders, like precision in general, shows only a moderate correlation with success ($r = .4$). The more successful predictors are somewhat less likely to be guilty of extreme errors, but are not free of them by any means.

Individual predictors, therefore, can be evaluated in terms of success, definiteness, and precision—how often they are right, how definitely they commit themselves to a prediction, and how accurately they estimate the probabilities of the given alternatives. While these aspects of predictive ability seem to be empirically associated to some degree, the results of the pilot study do not appear to indicate that they can all be regarded as stemming from some single skill in prediction.

3. Timing

The success of prediction presumably varies inversely with its scope in time—the relatively near future, that is, might be expected to be more predictable than the relatively distant future. This expectation is borne out by our data. What is most interesting, of course, is the *rate* at which reliability decreases with time. The results are summarized in Table 1.

Events that occurred within the first third of the time interval were successfully predicted in 49 per cent of all the predictions; in 45 per cent if they oc-

[7] Strictly speaking, these may not constitute blunders at all, since events of low probability may very well occur a few times and events of high probability fail to occur now and then. Blunders in estimates of probability can be judged only in the long run, with any degree of confidence.

curred in the second interval; and in only 35 per cent of the cases if they were most distant in time. It is especially striking that although the frequency of high values assigned was substantially the same whether the event predicted was in the relatively near or more distant future, the nearest events (among those to which high values were assigned) were predicted with 58 per cent success, and the most distant with only 43 per cent.[8]

Questions not verified until the full 20-week period of the study were predicted with even greater success than those of the nearest time group. This can be attributed, however, to the fact that these questions were verified, for the most part, without the occurrence of a *specific* verifying event. Twenty weeks could be recognized by most of the predictors as too short a time for the occurrence of events of the type asked about, and so a fair measure of success was achieved by selecting the "No Change" alternative.

This is shown by the following figures. All questions were divided in the analysis into two categories—"status quo" and "event-occurred" questions. The status quo (SQ) category included all questions for which the situation at the time of verification was identical, in relevant respects, with the situation at the time the question was asked; all other questions were put in the event-occurred category (EO).[9] Of all predictions on SQ questions, 59 per cent were successful, as against only 40 per cent on the EO questions. Now among the 84 questions in the 20-week group, 70 were in the SQ category, as compared with only 6 SQ among the 39 questions verified in less than 20 weeks. The high proportion of SQ questions in the 20-week group, together with the success achieved in SQ, thus may be taken to account for the apparent success in predicting the most distant events.

It is to be emphasized that this analysis of the time variable in prediction distinguishes between the relatively near and distant future *within* the near future from a sociological standpoint. Within a range of about five months, that is, nearer events are easier to predict, according to our results, than more distant ones. But nothing can be inferred from this data as to the relative predictability of events at a distance of five months and, say, five years. It is possible that, on this large time scale, the order of predictability is the reverse of what it is within small intervals.

4. Subject Matters

An important question in evaluating predictions is whether some types of events are easier to predict than others. In particular, the pilot study considered whether developments in science and technology are as predictable as those

[8] An inspection of the ten questions verified within one month, with 57 per cent of all the predictions being successful, supports the conclusions as to effect of time interval, at least to the extent of failing to reveal any striking features that might have made them especially easy to predict. Two questions were on the outcome of state primaries, and one each on rent control, tax legislation, a strike settlement, a resignation from a Cabinet office, Palestine casualties, a political assassination, an astronomical observation, and expected rainfall.

[9] If the question concerned an index (e.g., cost of living), it was classified as event-occurred even though the numerical value of the index remained unchanged; on the other hand, if a record (e.g., a speed record) remained unbroken, this was classified status quo. The criterion was whether the numerical measure was applied to identical or diverse events.

in political and economic affairs. It is to be noted that our predictors might have been expected to incline somewhat toward success on the former subject matters. The group consisted largely of persons trained in natural science—19 of the 26 fall in this broad category, and only 4 had special training in social science. Yet the few differences the results showed were in the direction of greater success in social science predictions.

Success in social science questions was 53 per cent as against 51 per cent in natural science. Predictions were made somewhat more confidently in the former case as well: values above 50 were used in 55 per cent of all social science predictions and in only 49 per cent of the others. The verifications in these higher values were also slightly more frequent for the social sciences: 66 per cent as against 64 per cent. For very high values (90 to 100), the difference was more marked: 77 per cent success in social science as compared with 67 per cent in natural science.

To be sure, most of these differences are too small to be significant, though they are always in the same direction. Many of the natural science questions, however, were successfully predicted because they were in the status quo category discussed above. Specifically, three-fourths of the natural science questions turned out to be SQ, but only one-half the social science questions. Hence, whatever differences there may be in the predictability of the two subject matters are masked by this factor. Among EO questions only, success in social science was 41 per cent, in natural science, 40 per cent. But for SQ questions, the figures were 64 per cent and 55 per cent, respectively. For high values (above 50), the differences are even more marked: 49 per cent to 41 per cent in EO, 78 per cent to 70 per cent in SQ—in both cases to the advantage of the social science subject matters. The results are summarized in Table 2.

Thus, though far from conclusive even for this sample, the data do suggest a somewhat greater predictability of the social as against natural science subject matters, especially when high values are used. It should be em-

TABLE 2

PREDICTABILITY OF SOCIAL SCIENCE AND NATURAL SCIENCE QUESTIONS

	Social Science Questions	*Natural Science Questions*
Number of questions	65	58
Number of predictions	1399	1254
Success per cent	53%	51%
Success per cent in Values above 50	66%	64%
Success per cent in Values above 89	77%	67%*
Success per cent in EO	41%	40%
Success per cent in SQ	64%	55%*
Success per cent in EO, Values above 50	49%	41%*
Success per cent in SQ, Values above 50	78%	70%*

* Statistically significant difference.

phasized that what is being compared here is not the ability to predict of the social and natural sciences, but rather the predictability of political and economic events as against developments in science and technology. Such developments are themselves, in large part, subject matter for the social sciences. Our results suggest, not that the sociologist can make better predictions than the physicist, but, what is very different, that it may be somewhat easier to predict political and economic phenomena than the results or rate of growth of science and technology.

It must also be recognized that the differences revealed in this study may be due in part, and perhaps altogether, to the fact that the predictors were not specialists on precisely the matters to be predicted. While it may be true that, for the non-specialist, social science subject matters are somewhat more predictable than those of the natural sciences, this need not hold for the specialist. Here again, the pilot study can only point out lines for further investigation.

B. THE PROBLEM OF IMPROVEMENT

1. The Effect of Knowledge

The simplest and most direct expedient that offers itself for the improvement of the success of predictions is the selection of only those made by the best informed predictors. Accordingly, some attempt was made in this study to examine the extent to which success in prediction can be ascribed to knowledge of the field in which the prediction is made. The 26 predictors were given the American Council on Education's "Cooperative General Culture Test" (Revised Series Form X), in two parts: "Current Social Problems," and "Science." Their scores, expressed as percentiles in terms of the results of college sophomores (in 1947), ranged from 5 to 100 in "Social Problems," the mean score of the group corresponding to a percentile of 75, and from 10 to 99 in "Science," with a mean corresponding to a percentile of 80.

These scores were then correlated with frequency of success in predictions in social science and natural science questions, respectively. The results were .6 in both cases. Predictive ability, of the type studied here, shows a significant positive correlation with knowledge as measured by these tests.[10]

The effect of such knowledge on success in this study is not, however, very great. The success of the best informed predictors was not vastly greater than that of the worst informed. In natural science, the top half of the group exhibited knowledge indicated by a mean percentile of 95.5, the mean percentile of the bottom half being 51.4. Yet the best informed group scored a mean success of only 55 per cent as compared with 49 per cent for the worst informed. For social science the results were similar: though the knowledge percentiles were 95.2 and 45, the success achieved was 57 per cent and 51 per cent, respectively.[11]

[10] The same is shown by frequency of blunders in the two fields, which also correlated (negatively) .3 and .6 with knowledge in social and natural science, respectively.

[11] Surprisingly small differences in success are shown even by the individual predictors occupying the extremes of knowledge. In natural science, the extremes in knowledge were percentiles of 99 and 10; the success scored by these two individuals was 56 and 48 per cent. In social science the corresponding figures are

It may be that the effect of knowledge was not shown to be greater because the tests used are an inadequate measure of the knowledge actually brought to bear in the predictions themselves. It may also be that there is operative in prediction a factor of "judgment" relatively independent of detailed knowledge. As to whether the first, second, or indeed some other explanation is preferable, our meager data on the psychological side of the matter provide no indication.

Whatever the bearing of knowledge on prediction, its effect could not be associated in this study with any particular group of questions. One might ask, that is, whether differences in predictive success can be attributed in part to differential skills in anticipating relatively unforeseeable events. We may envisage two possibilities: (1) The predictors attain an equal measure of success in easy predictions, but some of them are markedly more successful in the difficult predictions, thus giving a better over-all performance. (2) The more successful predictors are no better, relatively, in difficult than in easy predictions; their better performance is manifested equally in both.

The data definitely indicate a conclusion in favor of the second alternative. Questions were grouped according to the proportion of the whole set of predictors who answered them successfully, as a measure of their difficulty for the set of predictors as a whole. It was then found that there was no significant difference between the individuals in their relative frequency of success in questions of varying degrees of difficulty.

2. Groups

A second method of improving the success of prediction that suggests itself is to have the prediction made by several predictors together. A special phase of the study was directed towards answering this question. One quartet of predictors, the cooperative group, discussed the questions together and then answered them individually; another quartet, the joint group, discussed them until they reached a single collective decision. The joint group attained a success of 67 per cent and the cooperative group of 62 per cent. These figures are to be compared with the 52 per cent mean success of *the same individuals* when not participating in a group prediction. (See Table 3.) The group effort is thus significantly better than that of the individuals composing the group working independently.[12]

The question arises whether the higher success of the group is attributable to specific effects of collective effort, or whether a certain averaging takes place that could equally be obtained by statistical combination of individual results. The data strongly indicates that the latter is the case. Defining the *mean prediction* of a set of persons as that alternative to which the set assigned the highest mean value, we find that the mean prediction of the independent group attained a success—63 per cent—comparable to that of the cooperative and joint groups. The mean prediction of the cooperative group

percentiles of 100 and 5, and success of 55.5 per cent and 44 per cent.

[12] The difference in success of the joint and cooperative groups is not statistically significant.

TABLE 3

IMPROVEMENT OF PREDICTION

Method	*Success* (*Per Cent*)
All Predictors	53%
Best Informed Predictors (top half)	56
Worst Informed Predictors (bottom half)	50
Independent Group	52
Cooperative Group	62
Joint Group	67
Mean Prediction	66
Plurality Prediction	68
Best Individual Predictor	71

itself rose to only 63 per cent (from 62 per cent). For the entire set of 26 predictors, the success of the mean prediction was 66 per cent, as compared with the 53 per cent mean success of the individual predictors.[13]

It may be noted that these results are duplicated by the *plurality prediction* of a set—the alternative preferred by most members of the set. The success of the plurality predictions of the group as a whole was 68 per cent. Half the group, chosen at random, also scored a plurality success of 68 per cent, and the other half, 65 per cent.[14]

In short, in this study the success of collective psychological effort was duplicated by statistical methods.[15]

C. THE PROBLEM OF APPRAISAL

1. Rank Order

The appraisal of a prediction may be expected to be improved in accuracy by a knowledge of the rank order of likelihood assigned to the predicted alternative among a given set of alternatives. Plainly, the first choice among the alternatives can be expected to be verified more often than the second choice, and that in turn more often than the third.[16]

This expectation is fully confirmed by our data. The success in each rank was 52, 21, 14 and 13 per cent, in that order. First choice predictions, that is,

[13] This figure is fairly stable in our sample; the success of the mean predictions of a random set of half the predictors was 65 per cent, that of the other half, 67 per cent.

[14] As might be expected, however, the better predictors profited less (in absolute per cents of success) from statistical combination than the worse ones. The best 13 predictors had a mean success of 61 per cent, which rose to 70 per cent for the plurality prediction and to 72 per cent for the mean prediction. The worst predictors improved from 46 per cent to 61 per cent for both statistical combinations.

[15] Attempts to arrive at a consensus prediction—a mean weighted by past performance of the predictors—did not succeed in raising success. Analysis was made of the performance of three quartets, consisting of the four best, the four worst, and the four median predictors, respectively. The mean prediction was consistently more successful than the mean of individual successes, but a consensus based on least squares did not yield any stable improvement over the mean prediction.

[16] Usually, of course, the term "prediction" is applied only to the first choice alternative. It is convenient, however, to regard all four alternatives as having been predicted, though with varying degrees of confidence or estimated likelihood (including zero).

were successful more than twice as often as second choice predictions. The third and fourth choices did not differ from each other significantly in frequency of verification.

The higher ranks necessarily fall in a higher range of values: the first rank can range from 25 to 100, the second from 0 to 50, the third from 0 to 33, and the last from 0 to 25 only. Hence it is not clear from the preceding figures alone whether it is the rank of an alternative or the numerical value assigned to it that is most closely related to reliability.

The pure rank effect might be judged as follows. Values from 25 to 50 could be assigned to more than one alternative, with a tie for first place. (To determine rank for purposes of computation of rank success, such ties were broken by a random process.) Hence we may compare the success achieved by these values when they were undisputed for top rank with their success when some other alternative was given the same value. The difference may be construed as a measure of the rank effect, when value is constant.

A total of 403 predictions were made with a tie for first rank, the values ranging from 30 to 50, inclusive (the weighted mean of the values was 44.1). The success achieved in these predictions was 31.7 per cent. When the same values (weighted mean 44.7) occurred in untied first ranks, of which there were 702 cases, their success was 44.8 per cent. Thus, in round numbers, the comparative success of rank 1 and rank 1.5 predictions, *the values used being the same in both*, was 45 per cent as against 32 per cent. This indicates the utility of rank alone in appraising the reliability of prediction.

2. Level of Likelihood

If numerical estimates of likelihood are made in a prediction, presumably these numbers could be used to appraise the prediction, quite apart from the consideration of the ranking of alternatives. (In the present study the ranking was not made separately, but was contained in the numerical values assigned each alternative.) Alternatives assigned values of 90 or 100, for instance, may be expected to be verified more frequently than those assigned values of, say, 40 or 50.

We may ask first how the values were distributed by the various predictors and in the four ranks. Values from 90 to 100 were used by one predictor in 51 per cent of the questions; another predictor used such high values only once in his 123 predictions. The mean frequency of occurrence of the values 90 to 100 was 21 per cent. The frequency with which values from 51 to 89 were used ranged from 10 to 58 per cent of each individual's predictions, around a mean of 35 per cent. About one-fifth of the predictions, that is, were made with high estimates of likelihood, and about one-third more with likelihoods exceeding fifty.

The mean values in the four ranks were 62, 24, 10, and 5, in that order (with standard deviations of 27, 14, 10 and 7, respectively).[17] Values in the top rank, that is, were markedly higher than in the other three; the last two

[17] This corresponds to an index of definiteness, as defined on p. 98, of 23. This is lower than the mean definiteness of 27 given there, because here we have the definiteness of the *mean* likelihoods, which takes into account a certain amount of compensation in the estimates of the various predictors.

ranks were very nearly indistinguishable. *On the whole,* in other words, one of the four alternatives was selected by the entire group as the predicted one; indecision between the two most likely alternatives was relatively infrequent (15 per cent of all the predictions).

Because of this distribution of levels into ranks, differences in the success of the various levels might be construed as due to the rank differences which they incorporate. However, the success of the various levels differs markedly from one another *even when all of them are in the first rank*, i.e., assigned to the most preferred alternative. Values in the interval 25-50, when they were top rank, were successful in 39 per cent of the predictions; in the interval 51 to 89, in 61 per cent; and 90 or above, in 73 per cent.[18]

Thus the rank and level of a prediction provide a useful basis of appraisal of its chances of being verified. The results are summarized in Table 4.

TABLE 4

SUCCESS BY RANK AND VALUE

Rank	*Value Interval*	*Success (Per Cent)*
1	90-100	73%
1	51- 89	61
1	30- 50	45
1.5	30- 50	32
2	0- 50	21
3	0- 33	14
4	0- 25	13

3. Basis Statements

One of the possible ways of appraising a prediction is the examination of what the predictor says was the ground on which his prediction was based. Of course, from the viewpoint of its specific content, the ground can be criticized only by another expert. But the possibility remains that certain general logical criteria could be applied for the purposes of appraisal of the prediction based on it. This possibility was checked in the study, with markedly positive results.

Each question asked of the predictors was followed by a space labeled "Basis for Your Judgment." No instructions, oral or written, were given as to how this space was to be filled. Of the total of 2,653 answers, 39 per cent were accompanied by a basis statement.

These statements were then classified, prior to the verifications of the predictions, into four categories. Reliability of the classification was tested on a random sample of 49 of the actual comments. Four independent coders agreed unanimously on 37 items, and three of the four agreed on an additional 10. Only two items in the sample were classified by an evenly split decision. No items in this test were put into more than two categories. Taking a weighted mean yields a 93 per cent

[18] From the standpoint of individual performance, however, the highest values were not always significantly more successful; this was true for only nine of the 26 predictors. For these nine, that is, predictions were more likely to be successful if made with values of 90 or more than with values of 51 to 89. For the other predictors, the higher value did not provide a reliable basis for expectation of greater success. It is to be noted that these nine were by no means the most successful predictors of the group: four were less successful than average, five more so. It may be, therefore, that two distinct factors are involved in predictive ability: success, and self-appraisal. An expert in both senses would be one who is not only often right, but also knows when he is most likely to be right. (Compare the discussion of precision on p. 98.)

TABLE 5

SUCCESS BY BASIS STATEMENTS

Basis Category	*Number of Predictions*	*Frequency (Per Cent)*	*Success (Per Cent)*
"Guess"	290	11%	40%
"Rationalization	181	7	48
"Special"	67	2	55
"Justification"	499	19	62
No comment	1,616	61	51
Total	2,653	100%	52%

agreement in the classification. The categories were:

(1) "Justification": a statement which provides some degree of logical warrant for the prediction. The criterion is not whether the basis is a good one in *fact* (this could be judged only by another expert), but only whether it is a logically sound basis.[19] Justifications consist in factual elaborations of details of both question and answer, appeals to evidence of specific empirical generalizations, hypotheses about motivations of those whose behavior is being predicted, analyses of the time required for the event in question to occur, etc.

(2) "Rationalization": a statement which purports to provide a logical basis for the prediction but does not actually do so. Rationalizations consist in mere repetitions of the prediction; in references to completely unspecified "evidence of past experience"; in appeals to what is "reasonable," "obvious," etc.; and in mere statements of belief, either of the self or of another person not involved in the question at issue.

(3) "Guess": an explicit statement of ignorance or reliance on a random process. Not classified as guesses were estimates, often described as "conservative guess" or "best guess," etc.; nor were statements classified as guesses if the term (or a synonym) was used in conjunction with other arguments.

(4) "Special": comments on the question itself, rather than on the prediction made—e.g., on the wording of the question, its value, meaning, etc.

An analysis was then made of predictive performance in each of these basis categories. The results are summarized in Table 5.

Predictions for which none or only "special" comments were made did not differ significantly in success from all those commented on. But, as might be expected, "justifications" were markedly more successful than the other categories, 62 per cent of the predictions in this category being verified, as compared with 48 per cent for "ration-

[19] For instance, if the prediction that Dewey would be nominated for President by the Republicans were based on the argument that defeated candidates were always renominated, this would be classified as a "justification." (Actual examples are given in Appendix B.) It is not in fact a good reason—being false—but it is logically sound. Only a historian, not a logician, can find fault with it. Of course, to call a statement a "justification" is not to say that it completely justifies the conclusion beyond any doubt, but only to classify it as providing evidence of some weight or other, if it be true.

alizations." It may be noted that frequency of use of the "justification" basis per individual showed a small but significant positive correlation (.4) with the degree of success of that individual. (There is no significant correlation between success and any of the other basis categories.)

Predictions stated to be "guesses" were successful in 40 per cent of the cases—significantly more often than the expectation from a mechanically random process. This might be due to three factors: (1) The antecedent probabilities of the four alternatives may be sufficiently extreme so that predictions based only on these probabilities could be expected to be successful more often than 25 per cent of the time. (2) Predictions might be alleged to be guesses even when they are not, as a technique of self-defense: no blame attaches to failure, and success is construed as crediting the predictor with being a "brilliant guesser." (3) Predictions which are consciously guesses may be drawing upon knowledge and judgment of which the predictor is quite unaware. The present study affords no grounds for conclusions as to whether and to what extent these three factors were operative.

Thus the results of the basis classification suggest another procedure for improving estimates of the reliability of given predictions. Analysis of basis by a non-expert—i.e., without reference to what is actually known of the subject matter—is capable of discriminating sub-populations of predictions differing from one another in success in the range from 40 per cent to 62 per cent.

CONCLUSION

As was indicated in the introduction, the most serious question raised by a study of prediction is whether the analysis is made on a statistically stable population. The difficulties are threefold: those concerning the group of predictors, those concerning the questions asked, and those concerning the procedure.

Although the predictors in the pilot study were for the most part highly skilled specialists, they were not experts in the specific fields with which the predictions were concerned. This is especially important because each predictor was required to answer all the questions. It may be seriously doubted whether some of the results of this study would also appear if the predictions were made only by experts in the appropriate fields. For instance, though it is possible, as we have just seen, to improve the estimate of the reliability of prediction by a content analysis of basis statements when these are made by non-experts in the particular field, it may not be possible when they are made by scientists thoroughly equipped to discuss the matter of the prediction. In the latter case, it may be that "rationalizations" and "guesses" would appear too infrequently for such gross categories to be useful in appraisal.

The questions asked also afford a doubtful basis for extrapolation of the results to other cases. In the first place, there were statistically significant differences in the predictability of the events concerning which inquiry was made in the thirteen questionnaires. Secondly, the topics selected depended on the judgment of the experimenters regarding such factors as likelihood of occurrence of a verifying event within the specified time interval and of the intrinsic difficulty of the prediction. In the third place, the choice of the specific alternatives also depended on the experimenters' judgment. The study provides no basis for estimating the effects of a greater or lesser number of alternatives, of a different grouping into alternatives of all the possibilities, and of allowing the predictor to specify alternatives himself. Similarly, there is no check on the effect on predictability of wording, order, or other devices of emphasis in the questionnaires, or of the background information given.

Among the most serious difficulties in the pilot study was the extreme limitation in time imposed for reasons of practicality on predicted events—20 weeks. This had the effect, first, of enhancing predictability on the basis of obvious forecasts of "No Change," and second, of tending to force the selection of questions concerning the more rapidly changing types of events. The first effect is apparent in the results. In particular, predictions of scientific and technological developments could not be studied at all adequately for this reason. Three-fourths of the questions in this category turned out to be "status quo"—the time limit was too short to allow for prediction of specific developments. The effect of the time limit on selection of questions cannot be determined.

Other aspects of the procedure, though useful in making the analysis, also limit the applicability of the results to other cases. For instance, it was required that estimates of likelihood be numerically expressed on a 101-point scale. At the same time, no research on the questions was permitted, and very little background information was provided. The basis of the judgment had to be stated in only two or three sentences. In all these respects, the conditions of the study differed markedly from what actually occurs when an expert is asked for his prediction. What effect these differences have on the success of prediction cannot be determined on the basis of the study itself.

APPENDIX A

Sample Questions

1. (Asked on January 12, 1948) The monthly average of the BLS Cost of Living Index for September 1947 was 164 (1935-39= 100). The Cost of Living Index monthly average for April 1948 will be to the nearest integer:

a. Less than 160. a. ....
b. 161-165. b. ....
c. 166-170. c. ....
d. More than 170. d. ....
100%

Basis of your judgment:

2. (Asked on January 12, 1948) Charged particles have now been accelerated up to energies of less than 300 Mev. (million electron volts). Plans are now under way to construct more powerful accelerators. By J-Day energies will be announced of:

a. Up to 300 Mev. a. ....
b. 301-500 Mev. b. ....
c. 501-1000 Mev. c. ....
d. Over 1000 Mev. d. ....
100%

Basis of your judgment:

APPENDIX A (continued)

3. (Asked on February 3, 1948) The Republican party at its convention June 20th will select as its presidential candidate:

a. Dewey. a. ....
b. Stassen. b. ....
c. Taft. c. ....
d. Other. d. ....
100%

Basis of your judgment:

4. (Asked on February 17, 1948) On June 10, 1946, Italians elected by proportional representation a Constituent Assembly of 556 members, comprising 104 Communists, 115 Socialists, and 207 Christian Democrats. On April 18, 1948, there will be an election for 557 Deputies. Of these, the combined number of Communists and Socialists will be:

a. Less than 150. a. ....
b. 150-219. b. ....
c. 220-290. c. ....
d. Over 290. d. ....
100%

Basis of your judgment:

5. (Asked on March 16, 1948) The Radio Manufacturers Association estimated that the television output reached a new monthly peak in February with a jump of nearly 170 per cent in 6 months. There are now 19 stations operating in 22 cities, 82 have construction permits in 51 cities, and 93 applications are being investigated with 64 more pending. By August 3, 1948 the total number of television stations operating in the U.S. will be:

a. 19 or fewer. a. ....
b. 20-25. b. ....
c. 26-30. c. ....
d. More than 30. d. ....
100%

Basis of your judgment:

6. (Asked on March 29, 1948) In view of the danger of war, there is a possibility that (1) production of automobiles for civilian use will be legally restricted to save steel and (2) one or more auto factories will be converted to military production. By August 17, 1948, there will occur:

a. (1) only. a. ....
b. (2) only. b. ....
c. (1) and (2). c. ....
d. Neither (1) nor (2). d. ....
100%

Basis of your judgment:

APPENDIX B

Sample Basis Comments

1. "Justification":

(1) "The present unrest in England plus the increased activity of the Conservatives make (a) most unlikely. The opposition to the Labor Government would make (b)'s chances very slim. However, Attlee has enough support for (c)."

(2) "Don't believe airlines will adopt procedures by J-Day as they are quite expensive for emergency operations as airlines do not plan blind landings. For emergency use I believe the ground operated procedure is better although more expensive."

2. "Rationalization":

(1) "Not knowing anything about the relative merits of the two procedures, and assuming that both are of equal merit, it seems reasonable to assume that some airlines will adopt one and some will use the other, while the remainder will continue with neither."

(2) "An awful lot of things have to fail before a rail strike really happens."

3. "Guess":

(1) "Intuition."

(2) "Don't know anything about FCC."

4. "Special":

(1) "Ambiguous question to me. Thought we had the ground-operator system at a couple of airports now. What does (d) mean exactly?"

(2) "Why don't you give the figure for March 1947, since the figures may be seasonal?"