LONG-RANGE FORECASTING
From Crystal Ball to Computer
Six

JUDGMENTAL
METHODS

Contents

Types of Judgmental Forecasts .................. 81
Intentions ........................................ 81
Opinions .......................................... 85
Errors in Judgment .............................. 85
Bias .................................................. 86
Anchoring ......................................... 87
Selecting Judges .................................. 90
Intentions .......................................... 90
Expert Opinion .................................... 91
Posing the Question ............................. 96
Presenting Information .......................... 100
Scaling .............................................. 105
Sensitive Issues ................................... 107

INTERMISSION

Obtaining the Forecasts .......................... 108
Surveys ............................................ 111
Delphi .............................................. 116
Traditional Meetings ............................ 120
Structured Meetings .............................. 121
Group Depth Interview ........................... 124
Role Playing ...................................... 124
Choosing the Method ............................. 132
Combined Forecasts .............................. 132
Assessing Uncertainty ............................ 138
Self-Assessments ................................ 138
Comparisons among Forecasts ................. 144
Summary ......................................... 147
What a piece of work is a man! How noble in reason! How infinite in
faculty! In form and moving how express and admirable! In action how
like an angel! In apprehension how like a god! The beauty of the world!
The paragon of animals!

William Shakespeare*

The capacity of the human mind for formulating and solving complex
problems is very small compared with the size of problems whose solution
is required for objectively rational behavior in the real world—or even
for a reasonable approximation to such objective rationality.

Herbert Simon (1957, p. 198)*

There is quite a contrast between Shakespeare’s and Simon’s view-
points. On almost any basis one would choose Shakespeare! He is more
poetic than Simon; he is more widely read; and his position is more
popular. The only thing going for Simon is that he is right.

Shakespeare is expressing our hopes—a vision of the world as we
want to see it. His view is one of the little lies that we tell ourselves.
It makes us feel important.

We act as if Shakespeare were correct. Most forecasts are made using
judgmental methods. The more important the forecast, the more likely
we are to use judgmental methods in reaching it.

Judgmental forecasts have been well-studied. Much is known about
the problems involved in using such forecasts, and much is known
about how to improve judgmental forecasting.

What a wonderful position for a person who is writing a book on
forecasting! To be able to talk about an area that is important and
where people are generally doing things wrong, and to be able to say
how to put things right!

This chapter first describes the two sources of judgmental forecasts—
opinions and intentions. Bias and other problems with judgmental
forecasting are then described. The next three sections are devoted to
methods for obtaining the forecast: the selection of judges, the wording
of the questions, and the ways of obtaining the forecasts from the
judges. The use of combined forecasts is discussed. Finally, there is a
section on how to assess uncertainty. An intermission is provided in
the middle of the chapter.

*These quotations were contributed by Paul Slovic.
TYPES OF JUDGMENTAL FORECASTS

Intentions are statements that people make about their planned behavior, or about the behavior of things they can control. "Opinions" refer to things outside the judge's control; in this book, opinions refers to forecasts about events over which the judge has little control.

Intentions

Intentions data are most useful to the extent that the conditions in Exhibit 6-1 can be met. A discussion of the items in this list is provided.

---

**Exhibit 6-1 CONDITIONS FAVORING USE OF INTENTIONS DATA**

1. Event is important
2. Responses can be obtained
3. Respondent has a plan
4. Respondent reports correctly
5. Respondent can fulfill plan
6. New information is unlikely to change plan

---

**Event Is Important.** The more important the event, the more likely it is that intentions will provide good predictions (e.g., Tobin, 1959; Murray, 1969; Friend and Thomas, 1970). For example, intentions are useful for predictions about events that have an important effect upon the judge, such as the choice of a marriage partner or an occupation. They are somewhat useful for predictions about fairly important events like the purchase of automobiles (e.g., McNeil and Stoterau, 1967) TVs (Payne, 1975), and a TV service (STAPEL 1968). Intentions are of minor value in forecasting unimportant events like the choice of butter or toothpaste (e.g., Bird and Ehrenberg, 1966).

The perception of importance is influenced by time; events that occur in the near future are perceived as more important. This implies that long-range intentions are useful only for events having a major impact on the respondent's life. Evidence supporting these statements has been found by numerous researchers. The following studies are typical:

Pratt (1968) found a relationship between importance and planning time. The estimated planning time for major consumer du-
Judgmental Methods

Rables (clothes washers, dryers, and refrigerators) averaged about 13 weeks, and that for small durables (fans, radios, skillets) was about 2 weeks.

Clawson (1971) found that a 3-month consumer intentions survey was more closely related to behavior than 6-, 12-, and 24-month surveys.

Responses Can be Obtained. It is important that the relevant decision makers can be identified and reached. Easier said than done! This is one reason why so much literature exists on this subject. The literature is organized around three major types of errors—sampling, nonresponse, and response errors. **Sampling errors** create problems in generalizing from the sample to the population; nonresponse errors create problems in generalizing from the respondents to the sample; and response errors create problems in generalizing from the response to the individual respondent. Exhibit 6-2 illustrates these three sources of error for intentions data, assuming each type of error to be independent of the others.

An example may help to clarify Exhibit 6-2. Assume that you must forecast the sales of automobiles in the United States. You decide to ask people to forecast their own behavior. A sample is selected from the U.S. population, and its members are asked whether they expect to purchase a car within the next 6 months. You are interested in reducing the total error. Sampling errors result if the sample is too small or if it was selected from a list that was not representative of the population of potential automobile buyers. Nonresponse errors occur if individuals in the sample cannot be located, or if they refuse to answer. Response errors occur if the respondents do not know whether they will purchase cars, or if they want to answer in a way that makes them look good. Response error is affected by conditions 3 through 6 in Exhibit 6-1.

Respondent Has Plan. Intentions are most useful when the respondent has a plan (e.g., Westoff, 1958). The extent to which this plan is formalized affects its value as a forecast. Legal contracts provide a high degree of commitment. Engagements are good predictors of marriages.
Planned expenditures for plant and equipment provide good predictions of actual expenditures [WIMSATT and WOODWARD, 1970]. Wicker provides further support:

Wicker (1971) found that the correlation between intentions to donate to a church and actual donations was .92. In this case the church asked members to sign pledges, and it sent out quarterly statements.

**Respondent Reports Plan Correctly.** The respondent may fail to report a plan correctly for any number of reasons: he may fail to understand the question; may feel that his response will be used against him; he may be unable to express his intentions clearly; the plan may reflect poorly on him (in which case he may lie); or may simply be unwilling to tell you about his plan ("It is confidential, you know"). For these reasons, intention and behavior often differ substantially:
LaPiere’s (1934) study is often cited in this respect. He and a Chinese couple visited 250 hotels and restaurants and were refused service just once. Yet when a questionnaire was sent to the same places asking whether Chinese customers were welcome, 92% of the respondents said “no.” Another example is provided in Doob and Gross (1968): intentions data and experimental evidence yielded different predictions in their study of automobile horn-honking.

**Respondent Has Power to Fulfill Plan.** Respondents are best able to fulfill their plans when their actions are independent of the environment. Often, for example in political negotiations or in labor-management relationships, respondents’ actions are influenced by the actions of others. Interaction does not rule out the possibility of an intentions study, but it does increase the likelihood of response error.

**New Information Is Unlikely to Change Plan.** The value of the intentions forecast is reduced to the extent that changes occur between the time of the survey and the time of action. This is an obvious problem for political forecasts because often it is unclear what new information will occur (Watergate was an example in politics), or what information will be accepted as relevant by the decision makers (e.g., Watergate occurred well before the Nixon–McGovern election of 1972, yet most voters regarded it as irrelevant. On the other hand, a *Playboy* interview of President Carter appeared to be relevant to voters in the 1976 Democratic primaries).

The above conditions do arise often, so intentions surveys play an important role in forecasting. The technology in this area is well developed. Given these conditions, valid forecasts can be obtained by proper assessment of intentions. KALWANI and SILK [1982] present evidence on the validity of intentions in marketing. Intentions have long been useful in politics and the technology for measuring intention has improved over time:

PERRY [1979], in his review of the typical error in political forecasting, showed gains in accuracy for forecasts of voting in U.S. national elections:
Errors in Judgment

These gains have been achieved despite increasing refusal rates.

Opinions

Opinions data are more general than intentions data because they are not limited to situations where the respondent has an impact. Opinions data are also simpler to obtain. Sampling error is of little concern; in fact, forecasts are often made by using only one judge. Nonresponse errors are of little concern; if one expert refuses to answer, you find another expert. But there are many problems with response errors.

Economists have argued at great length over the relative value of opinions data and intentions data. How do these arguments get started? If you are interested in the topic you may want to examine Adams (1965) and Juster (1969). More relevant questions for our purposes are when should we use intentions data and when should we use opinions data.

Intentions studies seem most relevant if the conditions in Exhibit 6-1 are met. Otherwise, opinions studies should be used. This implies that intentions studies are useful for short-range forecasts, but that opinions surveys can be used for short- or long-range forecasting. From a cost viewpoint, intentions studies are most relevant to cases where there are few “intenders” and where they are easily located. Sometimes, both intentions and opinions can be used [e.g., SEWALL, 1981].

Errors in Judgment

This section discusses problems that you and I have as judges. They reduce our effectiveness in judgmental forecasting. Unfortunately, awareness of the problems is not always sufficient to overcome them.

Studies from social psychology document our difficulty in learning from the experience of others. Of particular interest is the study by Nisbett and Borgida (1975), where subjects’ predictions of their own behavior were not influenced when they were given information on
Judgmental Methods

how others acted in the same situation (where they were told to harm other people). These studies suggest that information about problems that others have in forecasting will not have much effect on our beliefs about our own actions. (Even this finding is hard to believe about ourselves.)

What is to be done in the face of such damning evidence? Two things: suspend judgment about yourself, and, more importantly, develop and follow an explicit strategy for judgmental forecasting. Don’t trust your common sense!

The errors in judgment are organized into two sections: bias and anchoring. These are only some of the errors. For a more complete treatment, see HOGARTH [1980] or KAHNEMAN, SLOVIC, and TVERSKY [1982].

Bias

Although bias can arise from the researcher and from the situation, the most serious form of bias is caused by the judge. Judges have preconceived notions about the world, and these can influence their forecasts. One form of bias has been called “optimism.” Predictions by judges reflect not only what they think will happen but also what they hope will happen. A good discussion on the effects of this bias is provided by Simon (1969). The effects are widespread, as shown many years ago:

Hayes (1936) surveyed people two weeks before the 1932 U.S. Presidential election. Of male factory workers who intended to vote for Hoover, 84% predicted that he would win. Of those who intended to vote for Roosevelt, only 6% thought that Hoover would win.

In McGregor (1938), opinions were used to forecast whether King Edward VIII would announce plans to marry within a year. In addition to being asked whether the King would marry, judges were also asked whether they thought he should marry. Of those who thought that the King should marry, 80% thought that he would marry. Of those who were indifferent, 60% thought that he would marry. Of those opposed to the marriage, only 32% thought that he would marry.
Kidd (1970) found that engineers were optimistic in predicting the time required to overhaul electric generators. The estimated time needed to complete a project was usually about 60% of the actual time required, even though the estimates were made after the project was well underway.

An analysis of annual earnings forecasts by 50 companies (Cope-land and Marioni, 1972) found an overestimate of earnings that averaged 16%.

These examples involve cases where judges were biased, but they did not stand to gain personally. Imagine the effect of bias when judges also benefit personally; for example, the medical researcher who is paid by a drug company to test the efficacy of one of its drugs, or the salesperson who is asked to forecast sales during the next year. Typically the greater the judge’s involvement in the forecast situation, the greater the expected bias.

Anchoring

“There is no reason for any individual to have a computer in their home,” said Ken Olson, president of Digital Equipment Corporation at a convention of the World Future Society in Boston, 1977 [CERF and NAVASKY, 1984].

This example illustrates anchoring. Anchoring is the tendency to start with an answer when making a forecast. More specifically, the example illustrates the type of anchoring that has been called conservatism. “Conservatism,” as defined here, is the assumption that the future will look like the past; there will be no abrupt changes.

Conservatism leads to underprediction of the amount of change. (For example, EGGLETON [1982] found that judgmental forecasts were more conservative than extrapolation forecasts.) The effects of conservatism are often confounded with the effects of optimism. This combined effect of bias and conservatism is illustrated by Herbert Hoover, who, on October 29, 1929 (just before the start of the Great Depression), said the following: “The fundamental business of the country, that is, the production and distribution of commodities, is on a sound and prosperous basis” (quoted by Cantril, 1938). The problem has been noted long ago, as shown in the following examples.
Ogburn (1934) found that students at colleges of losing football teams had forecasted that their teams would lose by an average of 3 points. The actual defeats averaged 18 points. Students at the winning colleges had forecasted victory by an average of 6 points.

Hultgren (1955) examined quarterly forecasts of tons of freight shipped over railroads from 1927 to 1952. The forecasts were made by experts employed by the railroad shippers. Forecast errors ranged from being too low by 1.7%, to being too high by 40.5%. The errors on the high side represented a bias toward good business.

Modigliani and Sauerlender (1955) also examined the railroad shippers’ one-quarter forecasts, this time from 1927 to 1941. They found that, when shipments were rising, shippers underestimated by 5%. But when shipments were falling, they overestimated by about 16%. On the average for the period, their forecasts were high by 4%.

Evidence contrary to conservatism has been difficult to find. However, Dorn (1950) suggested that population forecasts have not been conservative. The predicted changes in population have generally been too large.

Tversky and Kahneman (1974) found that previous experience is not the only basis for anchoring. It can be created by the way the question is asked. Or the judge may jump to conclusions on the basis of early evidence. For example, Webster (1964) found that interviewers make judgments of prospective employees in the first 30 seconds of the employment interview and stick to them. The following examples, although extreme, illustrate how easily anchoring can occur and how misleading it can be:

Tversky and Kahneman (1974) asked subjects to predict the percentage of nations in the United Nations that were African. A
starting value was selected by spinning a wheel of fortune in the subject's presence. The subject was asked to revise this number upward or downward to obtain his answer. This information-free starting value had a strong influence on the estimate. Those starting with 10% made predictions averaging 25%. This contrasts with the prediction of 45% by those given a starting value of 65%. (The correct answer is given in Appendix F.)

Multiplication Problem

Subjects are affected by initial data and tend to jump to conclusions. In Tversky and Kahneman (1974), subjects were asked to make rapid and intuitive estimates of math problems. One group was given $8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. (Quickly, now, can you guess the product? ______). Another group was given $1 \times 2 \times 3 \times 4 \times 5 \times 6 \times 7 \times 8$. This is the same problem, of course, but the median estimate for those getting the ascending sequence was 512, whereas the median estimate for the descending sequence was 2250. (The correct answer is given in Appendix F.)

There are numerous other problems that resemble anchoring. People expect the world to adjust to their current view; thus most of us suffer from the gambler's fallacy (Jarvik, 1951). A consecutive run of, say, four heads in coin tossing will lead many people to expect a tail on the next toss. The future is expected to compensate for unfair deviations in one direction. The National Weather Service predicted a very hot summer in the eastern United States in 1977 because of the very cold winter that had just ended. The temperatures would "average out" over the year. Might this happen to businesspeople who experience four bad years or investors who have four bad investments? Here is a simple example of the gambler's fallacy (Tversky and Kahneman, 1971):

Eighth Graders IQ Problem

The mean IQ of the population of eighth graders in a city is known to be 100. You have selected a random sample of 50 children for a study of educational achievements. The first child tested has an IQ of 150. What do you expect the mean IQ to be for the whole sample? ________ (The answer is given in Appendix F.)
Carlson's (1967) review of economic forecasts finds evidence of optimism, conservatism, and the gambler's fallacy, all combined. AHELERS and LAKINISHOK [1983] also found evidence of optimism and conservatism in their analysis of Livingston’s survey of economic forecasts.

Other errors arise with judgmental methods, but bias and anchoring are among the most serious. In the next section, I turn from errors to solutions. The solutions are considered for each of three steps in the development of judgmental models: selecting the judges, posing the question, and obtaining the forecasts. Ideally, one should proceed in the manner illustrated by the arrows in Exhibit 6-3. Typically, however, one must go back to an earlier step. Problems with bias may occur in any of these steps. It is expected to be more serious when uncertainty is high. Much can be done to correct for bias. Anchoring is more difficult to control, but some steps can still be taken.

**Exhibit 6-3 STEPS IN THE DEVELOPMENT OF JUDGMENTAL MODELS**

| Select judges | Pose questions | Obtain forecasts |

**SELECTING JUDGES**

The problems of selecting judges vary depending on whether one is using intentions or opinions data. This section provides advice on the selection of judges for each type of data.

**Intentions**

Careful selection of judges for intentions studies can reduce sampling error; this helps in generalizing from the sample to the population. The solution is probability sampling; it is well known and has been used for over half a century. Because you probably know about probability sampling, and because there are numerous good sources, the details are not discussed here.

Do not trust your intuitive feelings. Instead, rely on probability samples. Tversky and Kahneman (1971) illustrate the dangers of intuitive statistics, even for trained statisticians. Test yourself on the “sample-size-of-ten” problem. Most people do poorly on this—but, as Casey Stengel said, “You could look it up!”
Significance of Follow-up Samples

Suppose you have run an experiment on 20 subjects and have obtained a significant result that confirms your theory ($p < .05$ with two-tailed test). You now plan to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group?

$p = \underline{\text{______}}$ (The answer is given in Appendix F.)

Expert Opinion

"I think there is a world market for about five computers." This remark is attributed to Thomas J. Watson, Chairman of the Board of IBM in 1943 [CERF and NAVASKY, 1984].

People are willing to pay heavily for expert advice about the future. Milton Friedman is consulted to tell us how the economy will change: stock analysts are paid large salaries to forecast the earnings of various companies; Jean Dixon will forecast anything; political experts command large fees to tell our leaders what the future holds; and Jimmy the Greek forecasts the outcome of sporting events. The evidence says that this money is poorly spent. Because few people pay attention to results in this area, we might call this the "seer-sucker theory"; which is, "No matter how much evidence exists that seers do not exist, seers will find suckers."

Expertise beyond a minimal level in the subject that is being forecast is of little value in forecasting change. This conclusion represents one of the most surprising and useful findings in this chapter. It is surprising because emotionally, we cannot accept it. It is useful because the implication is obvious and clear cut: *Do not hire the best expert you can—or even close to the best. Hire the cheapest expert*. You have already recouped the cost of this book if you will follow this advice.

If this conclusion bothers you (it bothered me), you might try a short exercise so that the information in this chapter will be more useful. Take 30 minutes (time yourself), and describe on paper what information or evidence could possibly convince you that expertise is of no value in forecasting. This exercise requires a lot of psychic energy. If you wind up with a blank page after 30 minutes, you have just finished this section, in which case you can branch to the next section, "Posing the Question," on page 96. If you can list possible evidence, then you might find something helpful in this chapter.
Before proceeding with the evidence, let me make one thing perfectly clear: I am not against experts. No one who has read about Lem Putt, the privy builder (Sale, 1930), could be against experts. Some of my best friends are experts. If they keep their place in forecasting, everything will be fine. But their place is in saying how things are (estimating current status), rather than in predicting how things will be (forecasting change). The estimation of current status does, of course, play an important role in forecasting. (Expertise in forecasting methods is also valuable.)

Many studies have been done on the value of experts. Most have come from psychology and finance, but there is evidence also from economics, medicine, sports, and other areas. Expertise in the field of interest has been measured in various ways (education, experience, reputation, previous success, self-identification). Accuracy has also been measured in many ways. With few exceptions, the results fall into the pattern illustrated in Exhibit 6-4. Above the low level of expertise labeled $E_t$ (which can be obtained quickly and easily), expertise and accuracy are almost unrelated. It is likely, in fact, that accuracy drops off after expertise passes a certain level (indicated by the dotted line), but the evidence on this point is limited.

Evidence is available from well over 100 studies on the value of experts. Of these studies, only a few suggested that expertise improved forecast accuracy, and, even here, the gains were small. Of course, expertise does add to the comfort level of the client. A positive relationship would be expected between the client’s confidence and the money spent on experts.

Exhibit 6-4  RELATIONSHIP BETWEEN EXPERTISE AND ACCURACY IN FORECASTING CHANGE

![Graph showing the relationship between expertise and forecast accuracy](chart.png)
Selecting Judges

Psychology. Taft (1955) surveyed 81 studies from psychology that examined predictions made of others by experts and nonexperts. He concluded that nonpsychologists were more capable of making predictions about others than were psychology students or clinicians. The study was updated by Sarbin, Taft, and Bailey (1960), who added 10 studies. Grigg's study, described in the following example, is typical of the research that was reviewed, and more recent studies have been provided by Goldberg and by Levy and Ulman:

In Grigg (1958), 24 Ph.D.s, 24 trainees in psychology, and 24 naive subjects (undergraduates) listened to 10-minute interviews with each of three clients and then predicted how each client would fill out three different personality inventories. There was no difference in accuracy between the Ph.D.s and the trainees, and both were significantly more accurate than the naive subjects.

Goldberg (1965) found that 16 trainees did at least as well as 13 experts in diagnosing psychosis from scores on a personality inventory (the MMPI).

Levy and Ulman (1967) asked judges to distinguish 48 normal people from 48 psychiatric patients by an examination of paintings done by these subjects. All predictions were significantly better than chance, but expertise did not lead to significant gains in accuracy:

<table>
<thead>
<tr>
<th>Judges</th>
<th>Average Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Professional mental health workers</td>
<td>66.5</td>
</tr>
<tr>
<td>Student mental health workers</td>
<td>64.5</td>
</tr>
<tr>
<td>Persons with no mental health experience</td>
<td>64.6</td>
</tr>
<tr>
<td>Chance</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Additional studies in psychology by Grebstein (1963), Hiler and Nesvig (1965), and Oskamp (1967) reached the same conclusions.
Finance. Cox (1930) compared the performances of experts and novices in forecasting prices of stocks. He found no advantage for expertise. A study by Cowles followed shortly and reached the same conclusion, as have other studies:

Cowles (1933) examined 255 editorials by Hamilton, an editor of the *Wall Street Journal*, who had gained a reputation for successful forecasting. During the period from 1902 to 1929, Hamilton forecasted 90 changes in the market. Of these forecasts, 45 were correct, and 45 were incorrect. People following Hamilton's advice would not have done as well over this period as if they had made random investments. Cowles also found that a sample of 20 insurance companies did slightly worse than the market averages from 1928 to 1931, that 16 financial services did slightly worse than the average from 1928 to 1932, and that forecasts in 24 financial publications did slightly worse than the averages over this same period. A follow-up study (Cowles, 1944) reinforced these conclusions.

An excellent review of the literature in finance is provided by Granger and Morgenstern (1970). These studies (except for Shelton, 1967) add support to the conclusion that expertise is of little value in forecasting. More recent studies provided by Staël von Holstein (1972) and Richards (1976) yielded the same results.

An exception to the findings about the stock market is that expertise based on inside information is of value. (See Brooks (1969, Chapter 4), Lorie and Niederhoffer, (1968), and PENMAN [1982].)

Medicine, Sports, and Sociology. The first four studies described here suggest that expertise has no value. The next three suggest that expertise has minor value:

Johnston and McNeal (1967) had 12 judges predict the length of hospital stay for 379 mental patients over an 18-month period. Forecasts were made for 0 to 3 months vs. 4 months or longer. The judges' scores ranged from 63% to 86% correct. The more experienced judges were no more accurate.
Winkler (1971) examined forecasts for collegiate and professional football games for 1966 using 45 judges, and for 1967 using 10 judges. He found that expertise (as identified by the previous week's success or by self-ratings) was of little value.

AVISON and NETTLER [1976] found that expertise, measured by schooling, did not improve individual predictive accuracy for 52 items predicted in nine public opinion polls in the United States and Canada ($r^2 = .004$).

Armstrong and Overton (1977) found that two students and a housewife were as accurate in predicting which mail survey items would be subject to nonresponse bias as were nine faculty members who had substantial experience with survey research.

Kaplan, Skogstad, and Girshick (1950) found a small but statistically significant correlation for a sample of 26 forecasters between their accuracy in forecasting for 123 events in the social and natural sciences and their respective scores on items involving current social problems and science.

Winkler (1967) found that sportswriters did a little better than graduate students and faculty members on forecasting scores for collegiate and professional football games. The bookmakers' forecasts, in turn, were slightly better than the sportswriters'.

Wise (1976) examined 1556 predictions made publicly in the United States between 1890 and 1940. The predictions related to social, technological, economic, and political changes. Some evidence was found to support expertise, although the effects were small. The significance of the findings rested on one judge who could have been classified in either group.
As noted, these results are disturbing. They do not make much sense. Perhaps future research will offer new insight. However, given the evidence to date, I suggest that you avoid hiring the most expensive expert to forecast change. Instead, spread the budget over a number of less expensive experts. Also remember that experts are good at diagnosis (estimating the current status) and that is useful in forecasting.

Additional advice follows from the earlier discussion on bias: avoid using biased experts. Do not use anyone who will be affected by the forecast, and find experts who differ greatly on any factors believed to be related to the forecast. Then use more than one judge. How many? The gain from adding judges drops rapidly, while the cost per judge is fairly constant. Huber and Delbecq (1972) suggest using at least 5 judges; an increase to 10 gives better results, but thereafter improvements are minor. HOGARTH [1978], using theoretical arguments, concluded that you should use at least six but no more than 20 experts. LIBBY and BLASHFIELD [1978], using empirical studies, concluded that the optimum number of experts is between 5 and 9.

The argument is often made that we must tolerate some bias in our judges because their expertise is so great. For this reason, salespeople, politicians, stockbrokers, and others get involved in the forecast. Another explanation for our use of experts is that, by using them, we can avoid responsibility. For example, courts use expert ratings of the extent to which criminals are dangerous even though these ratings lack predictive validity [COCOZZA and STEADMAN, 1978].

POsing THE QUESTION

The wording of questions may have dramatic effects, as illustrated by this example from Hauser (1975); these questions were asked in 1940:

<table>
<thead>
<tr>
<th>Question</th>
<th>Percent Answering “Yes”</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. “Do you believe in freedom of speech?”</td>
<td>96</td>
</tr>
<tr>
<td>2. “Do you believe in freedom of speech to the extent of allowing radicals to hold meetings and express their views to the community?”</td>
<td>22</td>
</tr>
</tbody>
</table>
Incidentally, although this example occurred during wartime, a follow-up in 1970 again found a minority answering “yes” to question 2.

Here is a fine opportunity to use some of that good advice on research strategies from Chapter 4:

Tom W. Problem

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

The preceding personality sketch of Tom W. was written during Tom’s senior year in high school by a psychologist, on the basis of projective tests. Tom W. is currently a graduate student. Please rank the following nine fields of graduate specialization in order of the likelihood that Tom W. is a student in that field. Let rank 1 be the most probable choice.

_____ Business administration
_____ Computer science
_____ Engineering
_____ Humanities and education
_____ Law
_____ Library sciences
_____ Medicine
_____ Physical and life sciences
_____ Social science and social work

(Why not try your hand at the above problem before proceeding?)

This Tom W. prediction problem is borrowed from Kahneman and Tversky (1973). It is not a trick case, but it does illustrate some typical difficulties. In this study judges did not use all of their information in an efficient manner. They ranked the programs almost solely on the basis of the perceived similarity between the brief description given and a typical student in each program. The base rates for these graduate programs, that is, the fact that some programs are large relative to others, had almost no influence upon their predictions. As students
in psychology, most of these judges should have been aware that the personality sketch had no validity. In addition, all of these judges had been exposed to the notion of base-rate prediction in their statistical training. In short, the question, as stated, led the judges to make poor use of their knowledge.

What might be done to improve prediction in the Tom W. case? Consider the following possibilities:

1. **Decomposition** suggests that we identify the key components of the problem. In this case the two most obvious components would be the base rate and the individual variation from the base rate.

2. **Prior theory** can be used along with decomposition. In this case some judges would state that thumbnail personality sketches cannot be used to predict individual variation. They would then be left with the problem of predicting by using the base rate only.

3. **Causality** can be used in the statement of the problem to show why the base rate is relevant [AJZEN, 1977].

4. **Eclectic research** might be used by wording the question in different ways. One could vary the order of the answer categories, provide one version with the thumbnail sketch and one without it, rewrite the thumbnail sketch to eliminate evaluative wording, and use judges with different educational backgrounds. A combination of these forecasts would be expected to be more accurate.

For further comments on Tom W., see Appendix F.

The base rate problem has been the subject of much study. See AJZEN [1977], Carroll and Siegler (1977), CHRISTENSEN-SZALANSKI and BEACH [1982] and HOGARTH [1980]. For a practical application, see JOHNSON [1983].

These suggestions on wording can be applied to many questions on forecasting. In particular, one should distinguish between a judge’s knowledge of the current situation and his forecast of change (or between base rate and individual variation). For example, assume that you want to predict the number of suicides in 1985. No matter how you phrase the question, it will mean different things to different peo-
Posing the Question

people. Also, how much information should the question provide? A reasonable strategy is to word the question in different ways. If the current year was 1972, here are some possible ways the question might have been worded:

1. In 1970, there were 11.6 suicides per 100,000 people in the United States. What do you think the rate will be in 1985? _____

2. In 1970, there were 11.6 suicides per 100,000 people in the United States. Although the rate has fluctuated since 1900 (e.g., it rose in the 1930s), there has been no long-term trend. What do you think the rate will be in 1985? _____

3. By what percentage do you think the suicide rate will change between 1970 and 1985?

   Up by _____ %
   No change _____
   Down by _____ %

4. In 1970, there were 11.6 reported suicides per 100,000 population. Do you think that this correctly represents the suicide rate?

   Yes _____
   No, it is high by _____ %
   No, it is low by _____ %

(Incidentally, the data were taken from the U.S. Bureau of the Census, Historical Statistics of the United States, 1975, p. 414.)

Numerous possibilities exist for the introduction of bias into questions. HARRIS [1973] presents some examples; in one situation subjects were asked: “How tall was the basketball player?” Other subjects were asked, for the same player: “How short was the basketball player?” The respective averages for the player were 79 and 69 inches. Payne (1951) provides practical advice on the wording of questions, Noelle-Neumann (1970) discusses problems in wording and some solutions and SUDMAN and BRADBURN [1983] use the research findings to develop guidelines.

In the following subsections, specific techniques are discussed for posing the question. These include techniques for presenting information to the judges and for scaling the information received from the judges. These techniques are relevant to both opinions and intentions studies. Finally, consideration is given to questions for sensitive issues.
Judgmental Methods

Presenting Information

Cardinal Krol, Archbishop of Philadelphia, was asked about his position on the Vietnam War shortly after a visit to President Nixon in January 1972. A spokesperson answered something to the effect that "although the Cardinal was against war, he really could not make any judgment about Vietnam because he did not have all of the facts." Of course, President Nixon did not have "all of the facts" either. Presumably, however, Nixon did have more facts than Krol. Was this information likely to help Nixon to make a better judgment about the value of the war in Vietnam?

The example of Cardinal Krol illustrates one of man's basic strategies: he blames his difficulties on a lack of information. Thus, it is commonly assumed that the accuracy of a judge can be improved by giving him more information.

The research on the relationship between the amount of information provided to a judge and his predictive accuracy yields results that parallel those of the research on the value of expertise; that is, beyond a minimum amount of information, additional information does not add to accuracy—but it does add to cost. The familiar refrain "We need more information" is often incorrect. So here is another useful piece of advice: don't invest a great deal of money in obtaining better data. It is good practice to be suspicious when people tell you they need more information in order to make better predictions.

Most of the evidence on the value of information comes from psychology. Goldberg (1968a) reviewed evidence from 13 studies. Some typical examples of these studies are presented here, along with a study of a transportation problem:

Kelly and Fiske (1950) asked judges to predict success for participants in a training program in psychology. Information about the participants had been collected during 7 days of testing and interviewing. Predictions based on only a small portion of these data were at least as accurate as predictions based on all of the data.

Borke and Fiske (1957) asked clinical psychologists to make predictions about neurotic patients' responses to various paper-and-
pencil tests. Predictions based on a face-to-face interview were no better than those based on reading a verbatim transcript of an interview with these patients. Grigg (1958), in a similar study, reached the same conclusions.

Dudycha and Naylor (1966) created an artificial world, using a two-variable equation:

\[ Y = b_1X_1 + b_2X_2 \]

Evidence from this world (i.e., observations on \( X_1, X_2, \) and \( Y \)) were presented to subjects. Although both variables were valid, the addition of information on the less important variable did not improve accuracy in predicting, and in some cases it decreased the subjects accuracy.

Armstrong and Overton (1971) measured intentions to subscribe to a new type of urban transportation known as the Minicar system. A "brief" description, which was mailed to the sample, provided a picture of the car and described the service. A "comprehensive" description required that the subject attend a product clinic. Here the subject was guided through an exhibition of 18 wall graphs that explained key aspects of the service, was allowed to examine and to sit in the vehicle, and was shown a 14-minute movie that described the system. Guides were available to answer questions for the subjects. The cost per subject for the comprehensive description was about 10 times that for the brief description, not counting the cost of the prototype vehicle. Thirty-five subjects filled out the mail questionnaire on intentions to purchase the Minicar service. The same 35 subjects came in a week later for the comprehensive description. A comparison between intentions to purchase for the brief vs. the comprehensive description showed no difference. (Appendix F provides a brief description of the Minicar system.)

In summary, increasing the amount of information, that is, the number of bits of information, beyond a bare minimum, does not lead to improved forecasting by judges. (BRAUN and SRINIVASAN [1975]
provide clues as to the “bare minimum.”) This may be due to a limited ability to remember information (Miller, 1956a, b) or a limited ability to process information (Dudycha and Naylor, 1966).

The presentation of information is of greater concern than the amount of information. That is, how can a given amount of information be presented so that it is used most effectively by a judge?

Decomposition is one of the most effective ways to utilize an expert’s information. The expert can be asked to respond to questions on each of the parts of a problem. The analyst then synthesizes the responses to construct the forecast. Hertz (1964) illustrated this process using a capital budgeting decision. This strategy can be especially useful with a group of experts, each of whom may have information on only part of the problem.

Eclectic research can also help to utilize various bits of information. One might present different information to each expert, perhaps information tailored to utilize that person’s particular expertise. Responses from each of the judges would then be averaged. Pure speculation on my part.

The presentation of information on exponential growth is of particular interest, not only because it is common that things (like company sales) grow “percentagewise,” but also because this information is difficult for judges to handle. Studies by Wagenaar and Sagaria 1975 and Wagenaar and Timmers 1979 show that judgmental forecasts of exponential growth are highly conservative, partly because people tend to think in terms of unit differences rather than percentage changes. You can test this by asking someone to forecast the thickness of a thin piece of paper if it were folded in half 40 times. Show them the step by step folding up to eight folds, then ask them to forecast the thickness by judgment alone, assuming that it could be folded another 32 times. A group of 20 MBAs in my forecasting course at the University of Hawaii provided the following written predictions:

<table>
<thead>
<tr>
<th></th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than one foot</td>
<td>13</td>
</tr>
<tr>
<td>Greater than one foot, up to one mile</td>
<td>5</td>
</tr>
<tr>
<td>Greater than one mile, up to 2000 miles</td>
<td>2</td>
</tr>
<tr>
<td>Greater than 2000 miles</td>
<td>0</td>
</tr>
</tbody>
</table>

I have replicated the paper-folding exercise with other groups and found similar results. Few estimates exceed one mile. The actual answer is that it would be thick enough to stretch from the earth to the moon.
If judgmental forecasts must be made for exponential growth problems, two suggestions seem to help. You should:

1. Use fewer data; that is, do not examine the process too frequently [see WAGENAAR and TIMMERS, 1978; WAGENAAR, 1978] and
2. Present the historical growth as a decreasing function by using the inverse form, such as square miles per person rather than vice versa [see TIMMERS and WAGENAAR, 1977; WAGENAAR, 1978].

Even with these corrective procedures, the conservative bias often exists. Consider the following forecast of a decreasing function: “Where a calculator on the ENIAC is equipped with 18,000 vacuum tubes and weighs 30 tons, computers in the future may have only 1,000 vacuum tubes and perhaps weight 1 1/2 tons.” So said Popular Mechanics in March 1949. [CERF and NAVASKY, 1984].

For long-range forecasts, it is often necessary to forecast the net effect of a number of trends. For example, assume that you have asked experts to make predictions on the size and nature of the advertising industry in the year 2000. To help with this prediction you provide them with information on the expected ownership of video tape recorders by families, the use of cable TV, the capability of families to transmit messages on TV as well as to receive them, the capability of selling information via credit card and so on. Some people, myself included, have proposed the use of scenarios as an aid to forecasting because the scenario can integrate the information and help to show inter-relationships. Thus, you could write a story about a family in the year 2000 and show how they utilize all of this new technology. Promising?

In 1977, I contacted leading experts on scenario writing to assess its predictive validity. I was unable to find evidence either for or against the value of scenarios in forecasting. There were many applications, but no attempts to validate. On the other hand, the method seemed to have promise. For example, had I used a scenario, I would have been able to predict the usage rate for the rotary lawn sprinkler that I purchased. It would have gone like this: “Place sprinkler on lawn. Return to water faucet and turn on water. When lawn is soaked, go to water faucet and turn off water; then go to sprinkler and relocate; return to water faucet and turn on water.” In all, a frustrating and inefficient practice. As a result, the usage rate fell to zero. An alternative scenario using a sprinkler that oscillated would, I expect, have led me to predict a different usage rate.

As noted earlier, however (LRF pp. 40–45), the recent research on scenarios suggests that there are many pitfalls. Minor changes in the
way a scenario is written can have major impacts on the forecast of its likelihood. In light of this research, I do not recommend scenarios as a technique to improve accuracy.

When one tries to get "all the information," the likelihood increases of presenting irrelevant information to the judge. As we saw in the case of Tom W. (p. 97), judges often use irrelevant information, even when they realize its irrelevance. I have encountered the same situation among personnel recruiters who try to predict the future success of job candidates. Some recruiters have told me that grades are not a valid predictor of job success for an individual (this happens to be correct); but then they say that, as long as the candidates provide this information, they will use it! ZUKIER's [1982] experiment showed that the addition of irrelevant information will lead judges to make more conservative forecasts. ROOSE and DOHERTY [1976] and SLOVIC and McPHILLAMY [1974] found that this irrelevant information was especially likely to be used if it was "commensurate" (corresponding in measure) across observations; grading is a good example of an irrelevant yet commensurate measure for predicting success on the job.

Predictions for the Tom W. case can be greatly improved by omitting the irrelevant information. Kahneman and Tversky (1973, p. 241) tested this by asking subjects to predict for "Don":

"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, there was a strong relationship ($r = .74$) between the judge's estimate of the base rate and his prediction of field of study. The moral of the story is that no information is better than worthless information! One might even question the wisdom of providing any data at all to the experts. Objective data can be processed more effectively using an objective method. So why give them to the expert? Let the expert make the prediction on the basis of his own information; in this way information beyond the objective data may be utilized.

The strategy of letting the expert rely on his own information, however, is not without its difficulties. Consider the following problem:

"A word is selected at random from an English language dictionary. Is it more likely that the word starts with the letter $k$ or that $k$ is its third letter?"

[STOP AND CIRCLE ANSWER: 1st   3rd]
Posing the Question

To answer this, the subject searches his memory for relevant information. He will find that some information is easy to recall and some is difficult. In this case, it is easier to search for words by their first than by their third letters (try it and time yourself if you doubt this). Thus most people will recall more k-as-the-first-letter words and conclude that these are more common. In fact, k-as-the-third-letter words are three times as common. Tversky and Kahneman (1973) call this a problem of availability. Events that are easy to recall seem more common.

Scaling

It is difficult to avoid the problem of availability. My suggestion is that you ask the question in different ways. KRUGLANSKI, FRIEDLAND, and FARKASH (1984) showed that many biases can be avoided if the question is properly worded.

Here is a question that my state (Pennsylvania) asks school children in a survey called “Education Quality Assessment”:

“Someone whose skin is a different color from yours wants to borrow your coat:

(1) I would like it a lot.

(2) I would like it.

(3) I would not like it.

(4) I would hate it.”

This question is interesting because you are not allowed to be indifferent. As Pennsylvania’s question is worded, you can answer it as either a prejudiced or an unprejudiced person. (Your answer is assumed to relate to your feelings about race rather than to your feelings about your coat.) Don’t follow Pennsylvania’s example; instead allow for “no opinion.” In terms of scaling, this means that you should use scales with an odd number of points to handle such situations.

“Do you use 3-, 5-, or 7-point scales?” Now that is a popular question. It is asked so often you could hardly blame people for thinking that it is important. Actually, it is not very important, so this section is short. Much of the evidence is indirect and suggests that scaling is not important once you get past five categories, although small benefits accrue from more categories. Furthermore, there seems to be no danger in using more categories. Conclusion: use a lot of categories! Thomas Juster, who has done much research on this problem, suggested the use of 11-point scales for intentions studies. Huber and Delbecq (1972),
Judgmental Methods

who work with opinions studies, say that at least 5 categories should be used, and that an increase to 10 or more is better. COX [1980] reviews the research of this topic. Here is a description of Juster's work:

Juster (1964, 1966) compared an 11-point scale with a 3-point scale ("definite," "probable," "maybe"), in measuring intentions to purchase automobiles. Data were obtained from the 800 randomly selected respondents, the long scale being administered to them a few days after the short scale. Subsequent purchasing behavior of these respondents was then examined. The results indicated that the longer probability scale was able to explain about twice as much of the variance among the subsequent behavior of the judges as was the shorter scale. In addition, the mean value of the probability distribution for the 800 respondents on the 11-point scale provided a better estimate of the purchase rate for this group than the short scale.

If you care to do more reading on scaling, many sources exist. Consideration is given to the number of points on the scale, to the level of measurement, including nominal\textsuperscript{G}, ordinal\textsuperscript{G}, interval\textsuperscript{G}, and ratio\textsuperscript{G} scales (Stevens 1959), to the reference points for a scale (Rothman, 1964), and to other issues. Green and Tull (1978) and Selltiz, Wrightsman, and Cook (1976, Chapter 12) provide a general review of these issues.

When people are asked to predict the probabilities that each of a list of mutually exclusive and totally exhaustive events will occur, you might expect the total probabilities across events to sum to 100%. Certainly these probabilities should not sum to more than 100%! But many people do make predictions that sum to more than 100%, as shown by the four experiments reported in TEIGEN [1983]. Furthermore, the greater the number of possibilities, the greater the total probabilities. Furthermore, when two new possibilities were added to an original list of four, two-thirds of the subjects assigned probabilities without changing the probabilities for the original four alternatives, and only one-sixth of the subjects made adjustments that reduced the probabilities of the first four alternatives. The biases were substantial; for example, students in the social sciences predicted vocational choice for a youth seeking vocational guidance. When 10 possible choices were available, the total probability was 288%. Perhaps the solution is to
phrase questions so that respondents will not make such estimates. Or perhaps the responses can be normalized by the researcher so that they sum to 100%. Still another possibility is to frame the question as a series of two-way options. Further research is in order and, fortunately, this research can be replicated or extended at low cost. As a start, I tested the general principle by giving a question on marketing strategy to Wharton School MBA students. In a version of this question that offered three alternative courses of action, each of the 30 subjects assigned probabilities that totaled exactly 100%. When six alternatives were presented, five of the 26 respondents gave probabilities that summed to more than 100%, and the overall average was 105%. Although overestimation occurred, it was smaller than in Teigen’s experiments. The point of these studies is important: When soliciting probabilities, do not assume that the respondents will normalize their probabilities. (See also WRIGHT and WHALLEY [1983].

**Sensitive Issues**

"Would your friend do it?"

In some cases people may be reluctant to reveal their intentions. This typically occurs when the response may reflect poorly on the respondent. Sensitive issues include race, religion, politics, sex, and crime. For such issues one can use an indirect or projective test, that is, the judge responds to an ambiguous question or reports how someone else (e.g., a friend) would react. This technique removes the personal threat to the judge.

I was unable to find studies that compared the predictive validity of projective and direct methods. Different forecasts were implied, however, by the Nescafe study:

Haire (1950) asked women shoppers to describe the personalities of two women. The only information given about the women was their shopping lists. These lists were identical except that one included a regular coffee while the other included an instant coffee, Nescafe, which at the time was a new product. The woman with Nescafe was described by almost half the respondents as being lazy and a poor planner, whereas few respondents felt this way about the other shopper. They said that the Nescafe shopper did not care as much about her family because she would not spend the time to make percolator coffee. The implied forecast
was that, if Nescafé advertised that instant coffee saves time, sales would decrease. Housewives already knew that Nescafé saved time, but they felt guilty using it. Direct questions about instant coffee yielded a much different picture in this case. Respondents said that they did not drink instant coffee because they did not like the taste.

Few situations require the use of projective tests. They should be used when a truthful answer might put the respondent in a bad light. An extensive discussion of projective methods is provided in Selltiz, Wrightsman, and Cook (1976, Chapter 10).

INTERMISSION: Take a long break. You owe it to yourself!

OBTAINING THE FORECASTS

Once upon a time (actually in the early 1900s), there lived a German horse named Clever Hans. Clever Hans could answer an amazing variety of questions. Because it was hard for people to believe that a horse could be so smart, a commission was appointed to find out why he could answer questions. The commission found, for example, that Clever Hans could answer questions from strangers when his trainer was out of sight; thus, the trainer was not providing cues. The investigation stumbled about for some time until it was discovered that the answer lay not in Clever Hans, but in the person who asked the question. Clever Hans was able to answer a question only when the person asking it knew the answer! Without realizing it, questioners were giving cues about the correct answers. Generally they would lean forward to get a better view of Hans's hoof, at which point the horse would start tapping. When the correct answer approached, the questioner would often show an almost imperceptible head movement (that he was not aware of). This was the cue for Hans to stop tapping. He might still sound like a clever horse to you. Be reassured, however; Oskar Pfungst, who was leading the research commission, was able to follow the same type of cues as had Clever Hans, and he did almost as well as the horse in some experiments. (The original study is reported in Pfungst, 1911. For a brief summary, see Rosenthal and Rosnow, 1969, pp. 197–199.)

The story of Clever Hans suggests that the researcher can play an
important part in judgmental forecasting without realizing it. She may unwillingly communicate additional information that affects the judgmental forecasts.

Since the time of Clever Hans, numerous studies have examined the relationship between the researcher and the subject. A review of this research is provided by Rosenthal and Jacobson (1978), who refer to this problem as the Pygmalion effect—"one treated like a lady acts like a lady." Two typical studies are summarized here:

In Rosenthal and Jacobson (1968), students in an elementary school were tested by researchers. The researchers then provided false information to the teachers. They said that some students were of high potential, and they predicted that these students would do well during the school year. Although the students had been randomly assigned to the group, those predicted to do well were found to do better than the others in the group over the rest of the year.

Peters (1971) described a study similar to Rosenthal and Jacobson's. Students were told by their teacher that blue-eyed children were superior to brown-eyed children. In the space of less than one day, the blue-eyed children were performing better than their brown-eyed classmates.

(Why do I find these results so surprising? After all, I have been a student, I have worked as a "blue-collar" worker, and I was a private in the U.S. Army. In these situations I learned that "one treated like a child acts like a child." The Pygmalion effect applied to me too.)

The Pygmalion effect, a type of self-fulfilling prophesy, has often been attributed to cooperative subjects. In other words, subjects respond to please the researcher. According to this argument, judges try to provide the forecasts that the researcher is looking for. Sigall, Aronson, and Van Hoose (1970) reviewed the evidence in this area and found nothing to support this cooperative subject theory. The key point, they said, is that the subject wants to look good in the eyes of the researcher. They ran an experiment contrasting the two hypotheses and found that "looking good" was important, but "cooperation" was not.

To speak about the subject alone in the Pygmalion effect is analogous to studying clapping by examining the right hand only. The subject's
responses are interpreted by the researcher. This, too, can be a source of bias. "Bright young people" in companies may do well or seem to do well, because their superiors view them as bright young people. Rosenthal and Jacobson (1968, pp. 37-41) also summarized research showing how the researcher's bias can affect results. A typical study is provided by Rosenthal and Fode:

A class in experimental psychology was asked to conduct experiments with rats (Rosenthal and Fode, 1963). Five rats were given to each of 12 different student groups. Half of the groups were told that their rats were "bright," and half were told that their rats were "dull." In fact, there were no differences between the two groups. The rats were required to run through a maze. The average number of correct responses by the gifted rats was 2.3, and that by the disadvantaged rats was 1.5 (p < .01). ("Now, you parents get your children into those gifted programs at school. Too bad about the other kids," said Scott, in a fit of sarcasm.)

This discussion suggests that the relationship between the judge and the researcher may be a major source of error in judgmental forecasting. Another source of error arises from the interaction among judges. Moore's (1921) study implies that these two types of interaction have large and roughly equal impact upon the typical judge's opinions. Interaction among judges is desirable when the forecasted behavior is influenced by the opinions of other people. Examples are the intended purchases of clothing, automobiles, furniture, or vacations. Another important situation exists when one's actions are interdependent with the actions of others, as in political decisions, union-management negotiations, and some buyer-seller relationships.

More often, interaction among judges is undesirable. Rather than report true intentions or opinions, judges may report falsely to gain approval by the group. The effects are shown dramatically in research carried out by Asch:

In a series of trials, subjects were asked to match the length of a line with one of three unequal lines (Asch, 1965). Each of the subjects in the room was asked to announce his choice on each trial. All of the subjects but one were confederates of the exper-
Obtaining the Forecasts

They had been instructed to give obviously wrong answers on some trials (all confederates giving the same wrong answers). The situation was arranged so that the naive subject responded last. Under such conditions, only about 25% of the naive subjects continued to give correct answers. The rest showed some conformity, and 30% of the subjects gave wrong answers over half of the time. These results have been replicated many times.

The results of the Asch experiment were obtained in a situation where the answers were obvious. The effects of interaction should also be considered when the task is not so clear-cut, as is generally true in forecasting problems. Sherif's autokinetic study illustrates the effect of such interaction when the situation itself provides no information.

Sherif (1936) placed subjects in a dark room. A pinpoint of light was presented, and subjects were asked to judge how far the light moved. Although the light was stationary, it appeared to move because the observer had no point of reference. (This is called the autokinetic effect.) The group exerted a powerful influence on each subject's judgment of the movement. The group's earlier opinion also led to anchoring when new members were added to the group.

One other effect of interaction among judges is that the stated intentions may affect the behavior of the judges. This can be an important consideration when the judge is involved in the future situation, or when she feels committed to the prediction (Bennett, 1955, provides evidence on this).

In view of the importance of "judge–researcher" and "judge–judge" interaction, the discussion on how to obtain the forecast has been organized along these dimensions. Techniques for obtaining forecasts are plotted in their approximate locations in Exhibit 6-5. The discussion that follows considers each of the techniques.

Surveys

Survey methods avoid interaction among judges. This is advantageous if the judge's responses would be biased by how others react to her
statements, or if the responses of others would bias her responses. The assumption is made that other judges influence the responses but not the actual behavior. When the actual behavior would be influenced, it is important to account for this interaction; in such a case surveys would not be appropriate.

Surveys also assume that the respondent has a good knowledge of the situation and how she would act in it. If this is not the case, surveys should not be used. An example is provided by Hofling et al.:

In Hofling et al. (1966), 33 graduate and student nurses were asked on a questionnaire what they would do if a doctor tele-
phoned and asked that a patient be given a medicine. They knew that it was a violation of hospital policy to order medication by phone. Also they were told that they did not know the doctor, that the medicine was not on the authorized list, and that the dosage was twice the maximum allowable amount shown on the pill box. In response, 94% of the nurses said that they would not administer the dosage. A field experiment was also carried out on a comparable group of 22 nurses; only 1 nurse refused to administer the drug.

The basic survey methods are mail, telephone, and personal interview. Combinations of these methods exist, as do other approaches such as computer-led interviews (e.g., Greist et al., 1973). These approaches differ primarily in the amount of interaction among judges. Thus mail questionnaires (more precisely, self-administered questionnaires) involve less researcher–judges interaction than do telephone or personal interviews.

The relative merits of the various methods are briefly considered here by examining them against five criteria: cost, speed of response, sampling error, nonresponse bias, and response bias. A more complete discussion and further references are given in Boyd and Westfall (1981). Details on how to conduct mail and telephone surveys are provided in DILLMAN [1978] and Erdos and Morgan (1983).

Cost. Mail and telephone surveys are substantially cheaper per respondent than personal interviews. In some cases the personal interview is 3 to 10 times as expensive as mail and telephone surveys. The personal interview is especially expensive when the sampling units are widely dispersed geographically.

Speed. The telephone is clearly the fastest survey method. Many organizations are equipped to have interviewers feed responses directly to the computer, so summaries can be obtained immediately. Mail surveys are the slowest method, generally requiring at least a month. Personal interviews, although substantially faster than mail, are not as fast as telephone surveys. Of course, speed is generally not an important criterion for long-range forecasting.

Sampling Error. Often, it is difficult to obtain a list of the desired sample. In most situations the advantage lies with the personal interview. Mail surveys suffer because people change addresses. The most
serious problems, however, lie with telephone surveys; in addition to the fact that people move, about 10% of the population do not have telephones, and about 20% (mostly people who are younger, have less education, and have lower incomes) do not list their telephone numbers (Glasser and Metzger, 1972, 1975). The percentage of unlisted numbers is higher, about 30%, in urban areas. For these reasons, the lists for telephone surveys represent only about 70% of the households for general consumer intentions studies. Random digit dialing can be used for reaching the unlisted numbers, but this increases costs substantially (Glasser and Metzger, 1972) and it is a rude way to contact people.

**Nonresponse Bias.** The personal interview suffers least from nonresponse; it is more difficult to refuse someone face to face. Call-backs can be scheduled to ensure a high probability of reaching the respondent (Boyd and Westfall, 1970; Dunkelberg and Day, 1973). Although expensive, responses of 90% can sometimes be achieved.

Mail surveys have a serious problem with nonresponse. For example, general surveys face the problem that almost 10% of the population are not literate. Fortunately, however, much research exists on ways to reduce the nonresponse rate (Scott, 1961; Kanuk and Berenson, 1975; Linsky, 1975; Pressley, 1976; Erdos and Morgan, 1983). Follow-up has been shown to be an effective way to improve response rates. Monetary incentives up to about one dollar (1985 dollars) have also been found to be effective. (Armstrong, 1975b, reviews this literature.) Other useful practices are to use first class postage stamps for outgoing and return mail [do not use business reply postage, ARMSTRONG and LUSK, 1985], to ask short and relevant questions at the beginning, and to enclose an interesting cover letter. Surprisingly, length of questionnaire has a negligible effect. DILLMAN (1978) provides other useful advice. With techniques like these, one generally expects more than a 50% response rate, and 80% has often been achieved.

It is also possible to estimate the nonresponse bias in mail surveys. This can be done by reaching almost all of a subsample of the nonrespondents with a combination of mail, telephone, and personal contacts. This subsample can be used as an estimate for the nonrespondents. Another approach is to extrapolate responses across two or more waves. Armstrong and Overton (1977) show how simple extrapolations can reduce nonresponse bias by half.

Telephone surveys suffer greatly from nonresponse. Although nonresponse is reduced substantially by callbacks, even after as many as four callbacks not-at-homes may be in the neighborhood of 10%. Furthermore, it is easy to say “no” by telephone. Although some studies have achieved refusal rates lower than 5% (e.g., Kegeles et al., 1969),
the samples were special. More commonly about 20% refuse, and 35% is not unusual (Hauck and Cox, 1974; Falthzik, 1972). Falthzik suggests when to call in order to reduce the total of not-at-homes and refusals (Monday through Thursday from 9:00 A.M. to 12 noon are the best times). Overall, response rates of 70% are good for telephone surveys.

The best way to reduce nonresponse bias, when one considers both cost and error, is to use eclectic research. One may start, for example, with a mail survey, follow it with three mail follow-ups, and then telephone the nonrespondents. Nonresponse bias can then be estimated by extrapolation and by intensive efforts to contact a small subsample of nonrespondents. Of course, some people still will not respond. Simon (1969) says one should not be annoyed at them; the pay for this job is low, and it is a bother. It is amazing that so many people do respond to surveys.

Response Bias. Although George Gallup argued that personal interviews are most accurate (Gallup, 1976), the research suggests that for most issues the responses from mail, telephone, and personal interviews are similar. Rogers (1976), in a long survey focused mainly on education, obtained similar responses from telephone and personal interviews. Colombotos (1969) and Hochstim (1967) got similar results from mail, telephone, and personal interviews in health surveys.

The responses frequently differ on sensitive issues, however, including issues where “looking good” is of concern or where people feel that their response may be used against them. Sudman and Bradburn (1974) provide a thorough analysis of the literature on factors causing such response bias. Corrective procedures are examined in the review by Kalton and Schuman [1982]. The following examples show how differences may arise on sensitive topics:

O'Dell (1962) used mail and personal interviews to study consumer behavior. On some questions there were sharp differences: 37% of the people interviewed used hair rinse, while 51% of the mail respondents did so; 17% of the interviewees borrowed money at the bank, compared with 42% of the mail respondents.

Wiseman (1972a) compared mail, telephone, and personal interviews. For most questions there were no differences. Differences
were found, however, on sensitive issues. For example, for a question on the use of birth control, 75% of Catholic mail respondents were in favor, versus 44% of Catholic respondents interviewed by telephone or in person.

So far, the discussion has considered similarities and differences. But which type of survey is most accurate? Wiseman’s study suggests that mail surveys have an advantage for sensitive issues. This was verified by Sudman and Bradburn (1974) in their review of 935 studies; self-administered questionnaires (mail) are best for threatening questions. Cannell and Fowler (1963) found mail responses superior to personal interviews in a survey on hospitalization (records were available on the true situation in this study).

The superiority of mail surveys for reducing response bias was expected because there is less researcher-judge interaction. There is also less possibility for cheating, which may be a problem with telephone and personal interviewers (Roth, 1966). The problems can often be controlled in interview studies by training the interviewers and by monitoring some of the interviews to guard against cheating. These procedures can lead to results that are highly replicable, as shown by the following study.

In McMillan and Assael (1968), two different marketing research firms, Chilton Research and National Analysts, were given identical questionnaires and sampling plans and were asked to conduct independent interview studies on attitudes toward transportation. The responses were remarkably similar.

Projective questions might also serve to improve the capability of the survey methods to forecast sensitive issues. As noted earlier (LRF p. 107), however, this is speculation.

The advantages and disadvantages of the various survey methods are summarized in Exhibit 6-7 (LRF p. 134). But first the other methods for obtaining forecasts are examined.

Delphi

The Delphi technique obtains opinions through a mail survey. It uses the anonymity of responses from such a survey, and it typically adds three features:
Obtaining the Forecasts

1. The respondents are experts in the subject area.
2. There is more than one round—that is, the experts are asked for their opinions on each question more than one time.
3. Controlled feedback is provided. Respondents are told about the group’s responses on the preceding round. On round 2 and later rounds, respondents with extreme answers are sometimes asked to provide reasons, and these reasons are summarized anonymously for the next round.

Few techniques have captured the imagination of forecasters in such a grand manner as Delphi. It is truly a triumph of modern marketing. Would people be as captivated if it had been called “iterative mail surveys of experts?” It is much better to be named after the oracle at Delphi.

Delphi has been used for many years. Gerstenfeld (1971) found that over 10% of the firms in his sample of Fortune’s 500 had used Delphi. McHale’s (1973) survey of organizations, institutional units, and individuals engaged in futures research found that Delphi was one of the most popular techniques used. Hayden (1970), in a survey of “65 progressive companies,” found 26% of them using Delphi; of these, 71% claimed that it was useful.

Delphi has been the subject of hundreds of journal articles and numerous books. Linstone and Turoff (1975) provide a comprehensive review of this literature, and so does Martino (1983). Many variations of Delphi are presented, but most use the basic format as was described. Johnson (1976) provides a description of how Delphi was used by Corning Glass Works for a long-range market forecast. For another application, see BASU and SCHROEDER [1977].

Despite the vast literature, comparatively little empirical work has been done on the value of the three key aspects of Delphi (i.e., experts, iterative procedure, and feedback) over the typical mail survey. There have been many claims. For example, Martino (1983, pp. 26–29) stated that the selection of experts is the most important decision when using Delphi. The evidence on these three issues is examined in the following subsections.

Experts. The evidence reviewed early in this chapter stated that expertise in the subject is of little value in forecasting change. Why should it be any different for Delphi? Evidence in favor of experts in Delphi studies is lacking. The studies that have been published suggest high expertise is not necessary. Sackman (1975) refers to a study of computers that yielded similar results when replicated with graduate students instead of experts. Welty (1974) summarized additional evidence
Judgmental Methods

from five studies which showed that nonexperts can be used instead of experts; one of these studies is described here:

Welty (1972) replicated a Delphi forecast of American culture in the year 2000. The initial study had been done by Rescher using 58 experts. Welty used 43 college students majoring in sociology. Of the 17 items examined by Welty, there were no significant differences on 14.

On the other hand, expertise is expected to be helpful when Delphi is used to assess current status. This was supported by the following studies:

In Best (1974), 14 self-rated experts did better than 14 nonexperts in a Delphi study involving two questions involving estimates of current status. Similar results were found in a two-question Delphi study by Jolson and Rossow (1971).

Rounds. Certainly additional rounds produce greater agreement among Delphi experts. Some people think this is an argument in favor of adding rounds. Not so! The actual situation is just what you would expect. Sackman (1975) cites the similarity between results from autokinetic studies and from Delphi studies. The critical question is whether the average response moves toward greater accuracy on successive rounds. The answer is yes; there are small increases in accuracy, but it is not clear whether the gains are larger than one would obtain by adding experts rather than adding rounds.

Hample and Hilpert (1975) reanalyzed results from four Delphi studies. These studies, such as Dalkey (1969), used almanac questions (i.e., questions with known answers, such as "How many kangaroos are there in Australia?") whose answers were unlikely to be known by the panel. They found 214 questions where the median response changed on subsequent rounds. Improvements in accuracy were found for 63% of these questions. Additional studies, not included in Hample and Hilpert's analysis, include
Jolson and Rossow (1971), who found increases in accuracy on six of seven comparisons over three rounds, Best (1974), who found improvements on five of six comparisons over two rounds, PAR-ENTE, ANDERSON, and MYERS [1984] who found that additional rounds helped to predict when but not if an event would occur, and BOJE and MURNIGHAN [1982] who found no gain in accuracy over three rounds using almanac questions. Overall, it seems that additional rounds yield small gains in accuracy.

Feedback. Does it help to have people explain why they gave extreme answers? Hample and Hilpert (1975) examined this issue. Using data from two previous studies, they found that with no feedback, subsequent rounds were more accurate on 47% of the comparisons \((n = 30)\); with feedback, subsequent rounds were more accurate on 58% of the comparisons \((n = 45)\). Best (1974) reported a slight gain when feedback was provided for his two-question study. Gustafson et al. (1973) found no gain in accuracy; in fact, written feedback seemed to reduce accuracy. This limited evidence suggests that feedback produces slight gains in accuracy.

The Verdict on Delphi. This review suggests that high expertise is not required when Delphi is used to forecast change. Additional rounds and feedback each seem to contribute to accuracy, although the gains are modest and the evidence is limited.

Sackman (1975) provided a thorough review of Delphi. His criteria were strict, and his conclusion (his p. 74) was extreme: He said that "the massive liabilities of Delphi, in principle and in practice, outweigh its highly doubtful assets." After reviewing the Delphi literature and Sackman's critique, I agree with some of Sackman's points. Delphi researchers have been lax in examining Delphi; seldom do they study its predictive value, and seldom do they provide sufficient disclosure of their methods. On the other hand, Delphi has the advantages of the mail survey. The additional rounds and the feedback are of some value, perhaps of more value than increasing the size of a mail survey.

The primary advantage of Delphi is that it is a technique that is acceptable to organizations. It sounds fancy, yet the users can understand it. Groups using Delphi are almost as satisfied as those using traditional group meetings (Van de Ven, 1974). Most importantly, Delphi is more accurate than traditional group meetings. RIGGS [1983] provides some evidence on the accuracy of Delphi.

If you do use Delphi, observe the guidelines for good practice in
judgmental forecasting. As shown in STEWART and GLANTZ [1985],
expert surveys sometimes violate good practice.

Traditional Meetings

Remember the horse business problem (LRF p. 31)? It was stated that
about half of the individuals who do this arithmetic problem get it
wrong. If that depressed you, consider this: when the problem is given
to groups, about half of the groups miss it (Thomas and Fink, 1961).

Although groups can't get the horse business problem right, they
want to forecast the most important things in our lives. The traditional
meeting is the most commonly used approach to forecasting. Surveys
to support this conclusion (e.g., see PoKempner and Bailey, 1970), but
you already know this. The traditional group meeting has many prob-
lems, not the least of which is the strong pressure to conform. But you
know this too. Why, then, are traditional group meetings so popular?
Well, people generally do not invest much energy in preparation, so
meetings are easy. They also satisfy social needs; participants report
high satisfaction with groups (e.g., Van de Ven, 1974; BOJE and
MURNIGHAN, 1982). Finally, they help to satisfy power needs. It is
nice that they do a good job on social and power needs because they
certainly are not of much help in obtaining accurate forecasts. The
following are some examples of research studies on the value of groups.
Although Timmons (1942) cited studies showing advantages to groups
in certain cases, most studies since then have shown little value for
groups. The study by Kaplan is especially relevant as it deals directly
with forecasting.

In Jenness (1932), subjects estimated the number of beans in a
jar. Group discussion led to agreement, but did not improve ac-
curacy.

Kaplan, Skogstad, and Girshick (1950) studied 26 people who
made over 3,000 separate forecasts of 16 events in the social and
natural sciences. They found that traditional discussion in a four-
person group followed by individual forecasts led to correct pre-
dictions in 62% of the cases. As a basis for comparison, judges
working alone were statistically averaged as four-person groups,
and they were correct 63% of the time. In short, the group discussion did not improve forecast accuracy.

In Campbell (1968), the group solution to an organizational problem was inferior to the individual solution.

One could cite many more such studies. The group meeting is expected to do poorly—and it does. But instead of beating a dead horse, let’s examine ways to make the group meeting effective in forecasting.

**Structured Meetings**

It is especially important to consider improvement in group meetings because most organizations will continue to do their forecasting this way. A number of techniques exist to control “judge–judge” and “researcher–judge” interaction. Two of these methods are considered in this section: developmental discussions and estimate-talk-estimate (E-T-E^G).

The developmental discussion was adapted from Maier and Maier (1957). With this technique, a **facilitator** is used rather than a leader. Instead of a group that works for its leader, one has a facilitator who works for the group. The facilitator:

1. Prepares for the meeting by decomposing the problem.
2. Provides an opportunity in the meeting for all members of the group to participate, and especially encourages the expression of minority opinions.
3. Avoids evaluation and helps the group to suspend evaluation.
4. Avoids introducing her own ideas to the group.

Rules 3 and 4 reduce researcher-judge interaction. Rules 2 and 3 reduce judge–judge interaction. Rule 1 uses the group’s knowledge in a systematic way.

Consider again the horse business problem (LRF pp. 31, 120). Although group work was ineffective, the addition of a facilitator (using rules 2, 3, and 4) led to correct answers by 84% of the individuals (Maier and Solem, 1952). This improvement was significant (p < .05). It was not due to superior knowledge on the part of the facilitators (only 44% of the leaders had initially gotten the correct answer). Rather,
it was due to their success in allowing minority opinions to exert a constructive influence.

Most of the research on developmental discussion deals with problem solving rather than with forecasting. The results support the advantages of developmental discussion. The study described below by Maier and Maier found decomposition to be useful, and Hall reported that better predictions were obtained if evaluation was suspended, and an opportunity was given to express minority opinions.

Maier and Maier (1957) compared discussion groups using rules 2, 3, and 4 with developmental discussion groups using all four rules in order to assess the value of decomposition. The problem was "the case of Viola Burns," an employee in an organization. Background information was provided, and groups were asked to decide whether Viola should be encouraged or discouraged from taking a new job. The case was designed so that the highest-quality judgment would be to discourage Viola from taking the job. Of the developmental discussion groups, 40% reached this decision vs. 18% for the free discussion groups.

Hall (1971) used the "lost-on-the-moon" problem. Participants were asked to predict which items in a given list experts would recommend as being necessary for a survival trip across the moon (see Appendix F for a full statement of the problem). After 148 upper management personnel from several business organizations had been randomly divided into 32 discussion groups of four to six members each, half of the groups were given instructions on how to reach a consensus. These instructions stressed the suspension of evaluation and promoted the expression of minority opinions (rules 2 and 3). These groups provided more accurate predictions than did groups not given instructions. HERBERT and YOST [1979] replicated this study and obtained similar results.

Hall, Mouton, and Blake (1963) asked subjects to predict the order in which jurors in the movie Twelve Angry Men would shift from a vote of "guilty" to "not guilty." Consensus rules led to improved
Obtaining the Forecasts

predictions for 17 of 22 groups ($p < .05$). Similar results were obtained for the same problem by Holloman and Hendrick (1972).

Another method for structuring meetings is E-T-E. This technique is similar to Delphi except that a face-to-face meeting is used instead of a mail survey. Estimates in E-T-E are made anonymously. During the talk period, participants are asked to avoid arguing for their own position or even revealing it. Unlike Delphi, however, the researcher does not generally control the discussion, and people will sometimes argue for their own positions.

In comparison to Delphi, E-T-E offers advantages in terms of speed of response. The process can be completed within a matter of minutes or hours rather than weeks or months (Van de Ven's 1974 Delphi study took five months, for example). Although people enjoy unstructured meetings a bit more and they think they are more effective than E-T-E [BOJE and MURNIGHAN, 1982], E-T-E provides greater accuracy than do traditional meetings:

Gustafson et al. (1973) found slightly more accurate forecasts for E-T-E than for traditional meetings in a study asking judges to predict on questions such as this: "The observed weight of a person is 130 pounds. Is the person more likely to be male or female?"

There is, however, a disturbing finding that perhaps meetings should be eliminated. In other words, people would simply record their predictions. Hardly a social event, but the predictions would probably be about as good, and time would be saved. ROHRBAUGH's [1979] study on predicting grade point averages for students, and BORMAN's [1982] study of personnel predictions supports such a conclusion, as did the following studies:

Dalkey (1967), using almanac questions, concluded that anonymous individual forecasts (carried out before a group meeting) were more accurate than those obtained after a discussion among the group members.
Campbell (1968) found that group discussion of an organizational problem led to small but insignificant improvements in the quality of the decisions in comparison with prior decisions by the individuals.

One variation of E-T-E is to have an individual and anonymous first estimate, which is followed by a group discussion and a group decision. Unfortunately, the group decision seems to be inferior to the original pooled estimate of the individuals (Campbell, 1968; Maier, 1973).

Van de Ven (1974) proposes a "nominal group meeting." This is a variation of E-T-E in which more control is exerted over the discussion. A related possibility is to use a developmental discussion during the talk phase. The effects of these procedures on forecasting have not been assessed.

The best advice is to use some structured technique. FISCHER (1981) found few differences in his comparisons of the accuracy of various structured techniques.

**Group Depth Interview**

The group depth interview is useful for intentions studies when the behavior of the judge will be influenced by the responses of others and when the respondent is unsure how others will respond. This might occur for decisions such as the use of kilts by men, the purchase of a microwave oven or an electric car, or the adoption of a new approach to education by a school. Group depth interviews are more relevant to the extent that the changes are viewed as important, approval by others is evident, and uncertainty is high.

The group depth interview is basically a nondirective interview (it follows the rules provided on LRF pp. 28–31). The differences are that one is dealing with a group of perhaps four to eight people, rather than with an individual, and the scope of the initial question is generally narrower. Because of this narrower focus, the technique is sometimes called a focused group interview. (See Goldman, 1962, for an alternative description of this technique.)

Although the group depth interview is reasonable and widely used, I have been unable to find studies on its value in forecasting problems.

**Role Playing**

During the Vietnam peace action, high-ranking officers in the U.S. military played a game. They split into two groups, one representing
the United States and the other representing North Vietnam. Various strategies were then examined. In particular, the bombing of North Vietnam was considered. As a strategy for the United States, limited bombing was a military failure in this role play. Unlimited bombing was a bit more favorable from a military viewpoint, but it had no clearcut advantage over no bombing at all. The decision makers ultimately ignored the results of this game and chose the poorest strategy from a military viewpoint: limited bombing. The prediction from the role playing was accurate; limited bombing was ineffective and costly (Halberstam, 1973).

In role playing, people act as if they were the actors in the situation to be forecast. The role play should be realistic. The following rules should be followed when role playing is used:

1. Use “props” to make the situation realistic. Special clothing sometimes helps; for example, in the study by Janis and Mann (1965) the person who played the role of the doctor was asked to wear a white coat. In some situations, it helps to rearrange the furniture.
2. The actors should not step out of their roles, that is, once the actors meet, they should “be” that person at all times. A good idea here is to ask the role players to separate, and then return to the meeting place a few minutes later, after they have mentally prepared themselves and are ready to stay with their new identities.
3. The actors should improvise as needed and should throw themselves into the role-playing session.

These rules can be implemented with a low budget. Elaborate simulations do not seem necessary [e.g., see ELSTEIN, SHULMAN, and SPAFKA, 1978; ARMSTRONG and WALKER 1983.1].

Some role-playing situations have been found to create a great deal of realism. In addition to the Vietnam case mentioned, one could cite Zimbardo’s (1972) simulation of a prison. The players in this case were surprised at their own behavior. The action was terminated prematurely because the realism seemed to be getting out of hand, and the safety of the “prisoners” was threatened. Orne, Sheehan, and Evans (1968) found that role playing was realistic enough that observers could not distinguish between subjects who were hypnotized and those who were role-playing a hypnotic trance.

When role playing is used to forecast, two possible instructions may be used. One is to ask the players “to act as you yourself would act in this situation.” The other is to ask them “to act as you think the person you are playing would act.” I have used the former, as have most of the studies cited in this section, but I am not sure about the effects of
these differing instructions. This approach assumes that the role is the dominant influence on behavior. The instruction to “act as that person would act” would incorporate that individual’s personal preferences. This instruction requires some knowledge about the person.

Role playing seems to be most valuable when the interaction among people has a lot to do with the responses of these people. In addition, role playing seems to be of value when:

1. There are conflicts among the people involved.
2. It is not feasible to obtain intentions data from the relevant people (e.g., the various people involved in the Falkland Islands crisis between Argentina and the United Kingdom).
3. The decision is an important one.

The role-playing approach has been popular in political science, where it is called gaming. Goldhamer and Speier (1959) report that Germany used role playing in 1929 to plan its war strategy; Japan did so in 1940; Russia has engaged in it periodically and Herman and Herman (1967) used it to simulate the outbreak of World War I.

The use of role playing in business seems to be limited. Busch (1961) said that the Lockheed Corporation would ask executives to stand in the shoes of their major customers during group discussions and that they found this to be a satisfactory way to forecast the behavior of these customers. Evidence from mock juries suggests that they provide valid predictions [KERR et al., 1979]. IBM used a form of role playing to predict the reactions of a jury in a trial (Wall Street Journal, Feb. 3, 1977, p. 7). BORMAN (1982) found role playing to yield accurate personnel predictions.

Predictions from role playing may differ greatly from those provided by other methods. Here is a simple example:

Cyert, March, and Starbuck (1961) divided judges into two groups of 16 each. Each group was assigned a different role. One role asked the judge to assume that as the chief cost analyst for a manufacturing concern, he was to produce a cost estimate on the basis of preliminary estimates provided by two assistants in whom he had equal confidence. In the other role, the judge was the chief market analyst and he had to provide a sales estimate. The data were identical for both roles. The cost analysts estimated on the high side, and the market analysts estimated on the low side. Seldom did the analyst simply average the estimates from his
two assistants—the expected behavior if no role had been assigned.

It is only in recent years that the predictive value of role playing has been studied. In some of these studies, role playing was used to predict results that had been obtained experimentally. I found 13 studies using this approach; of these, 11 provided results that could have been used to predict the results of the experiment. The following are typical studies:

Greenberg (1967) used role playing and successfully replicated a laboratory experiment showing that higher anxiety produces a greater need for affiliation in first-born children. The subjects were told, “It is very important that you take your role seriously and that you act as if this were a real situation.”

Willis and Willis (1970) told subjects about the purpose and design of a laboratory experiment on conformity and then asked them to play the role of naive subjects. The main effect agreed with those from the laboratory study, but no interaction effect was found.

Mixon (1972), using a crude role playing procedure, was able successfully to replicate the Milgram (1974) laboratory experiment on obedience. This is noteworthy because opinions surveys of psychiatrists and intentions surveys of potential participants in the experiment yielded predictions dramatically different from those obtained in the Milgram study (Milgram, 1974, pp. 27–31; Larsen et al., 1972).

Additional studies include Berscheid et al. (1973); Darroch and Steiner (1970); Holmes and Bennett (1974); Horowitz and Rothschild (1970); Houston and Holmes (1975); Ring, Wallston, and Corey (1970); Simons and Piliavin (1972); Terry (1974); Wexley,
Singh, and Yukl (1973); and Yinon, Shoham, and Lewis (1974). This has been a controversial topic in psychology. Viewpoints on the value of role playing may be found in Hamilton (1976), Freedman (1969), and Miller (1972).

The predictive value of role playing for real-life situations has been tested in eight situations. Overall, role playing yielded 70% correct predictions, in comparison with 21% correct predictions for opinions. Some of these studies are summarized here:

Crow and Noel (1965) had subjects role-play a historical situation. The case, known as the "East Algonian exercise," was based on the annexation of Texas by the United States. The authors stated that it was unlikely that their judges, who were students, could correctly identify the true situation. To test this, I asked John Cronin, a professor who specializes in military history at Delaware County Community College, in Pennsylvania, whether he could identify the situation. He spent a substantial amount of time but was unable to make a correct identification. The case was presented by Crow and Noel to 96 different groups. Each group was asked to role-play to reach, for the Mexican leader, a decision ranging from 1, a peaceful response, to 11, a warlike response. Although historians claim that the optimal decision would have been 1 or 2, only 1% of the groups reached these decisions. Fifty-seven percent reached 4 or 5. The actual decision was classified by historians as 4 or 5, and it proved disastrous for Mexico. In other words, the role playing yielded accurate predictions of the behavior of the Mexican leader in a problem where the prediction was not obvious.

Since 1908, Washington and Lee University has been running a mock political convention every four years to select a Presidential candidate for the party that is out of power. This convention is generally run in early May before the actual convention, which typically occurs in July or August. (A description of the details of this elaborate role play may be obtained from Washington and Lee University, Lexington, Virginia, 24450.) Including the Mondale nomination in 1984, the convention has been accurate on 13
of 18 candidates. In this case, however, public opinion polls, which have been conducted since 1936, provide a comparable record of accuracy (Runyon, Verdini, and Runyon, 1971; Gallup Opinion Index for 1972 and 1976; and the Harris survey for 1980 and 1984). The candidate who was leading in the Gallup poll at the same time as the Washington and Lee convention won on 8 of 12 occasions. During the same period, the role-playing prediction was also correct on 8 of 12 occasions. (The two approaches agreed on 7 occasions, and, when they did, all predictions were correct.)

Armstrong (1977) asked subjects to play the roles of seven members of the board of directors for the Upjohn Corporation. They were told that an unbiased group of medical scientists, after 20 years of study, were unanimously recommending that Panalba, an Upjohn drug with harmful side effects, be removed from the market. The board was given 45 minutes to agree on one of the following five decisions:

1. Recall Panalba immediately and destroy.
2. Stop production of Panalba immediately but allow what’s made to be sold.
3. Stop all advertising and promotion of Panalba, but provide it for those doctors that request it.
4. Continue efforts to most effectively market Panalba until sale is actually banned.
5. Continue efforts to most effectively market Panalba, and take legal, political, and other necessary actions to prevent the authorities from banning Panalba.

Of the 57 groups that played the role faced by Upjohn, none removed the drug from the market. Furthermore, 79% decided to take decision 5. In fact, Upjohn also took decision 5. (Don’t worry; the Supreme Court made the company take the drug off the market in 1970. On the other hand, if you travel in foreign countries, you still have a chance to purchase Panalba, but under a different name. Or so I am told; the Upjohn Corporation has been too busy to answer my letters. See Mintz, 1969, for descriptions of the events in this case.) The role-playing predictions in the Upjohn case differed substantially from the predictions by nonrole-play-
ers of what they would do in this situation. Only 2% of 71 respondents to an interview said they would select decision 5, and over half said they would choose decision 1 (to remove the drug from the market). In an opinions survey, 41% of the 46 respondents predicted that Upjohn would select decision 5.

ARMSTRONG and WALKER [1983] randomly assigned pairs of subjects (Wharton School MBA students) to make predictions by either giving opinions or by using role playing. Three situations were used, two being disguised versions of historical events, while the third was an event that had not yet occurred. In a test using the first historical event, subjects were asked to predict whether a supermarket chain would accept the proposal by a manufacturer to sell consumer durables in their stores. This was based on a plan developed by Philco (remember them?) in 1961. Subjects in the group giving their expert opinion regarded the plan as unrealistic and predicted that the supermarkets would not accept it. Those who role played the situation made substantially different predictions. In fact, the plan was accepted, and it proved to be a disaster. The second historical situation involved a sit-in by artists in the major art museum in Histavia. Subjects were asked to predict whether government leaders would give in to the demands by the artists that the government extend its program of purchasing artwork that the artists could not otherwise sell. This was based on an event in Holland as reported in the January 7, 1982 issue of the Wall Street Journal. The government did cave in, a response that was not predicted by opinions; the role players had a bit of success on this prediction. The event that had not yet occurred involved a possible strike by the National Football League players in 1982. Those subjects using expert opinion predicted no strike, while the role-playing groups predicted a strike. (As is well known to you football fans, there was a strike and it crippled the nation for many weeks.) One possible defect in the design was that the role descriptions contained additional relevant information. To test this, we ran a “role aware” version of the distribution case, whereby those who were to give expert opinions were provided with the roles along with the background information. This had no effect on the results. The following table summarizes our results:
### Role Playing vs. Opinions in Predicting Outcomes

(Entries represent predictions by pairs of subjects)

<table>
<thead>
<tr>
<th>Situation</th>
<th>Prediction Same as Actual?</th>
<th>Opinion Normal</th>
<th>Role Aware</th>
<th>Total</th>
<th>Role Play</th>
<th>Statistical Significance*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
<td>No</td>
<td>15</td>
<td>18</td>
<td>33</td>
<td>2</td>
<td>P &lt; .01</td>
</tr>
<tr>
<td>plan</td>
<td>Yes</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Artists'</td>
<td>No</td>
<td>13</td>
<td>0</td>
<td>13</td>
<td>2</td>
<td>P = 10</td>
</tr>
<tr>
<td>reprieve</td>
<td>Yes</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>No answer</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>NFL 55% plan</td>
<td>No</td>
<td>11</td>
<td>0</td>
<td>11</td>
<td>4</td>
<td>P = 10</td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>

*One-tailed Fisher Exact Test comparing opinions with role play.
A summary of the preceding evidence on actual situations provides much evidence favoring role playing. As shown in Exhibit 6-6, role playing was correct on 70% of the predictions versus 21% for expert opinions (the predictive accuracy of opinions was about the same as would be obtained by chance). Though the sample size is small, the evidence is consistent with that from the experimental studies by psychologists as noted. These gains in accuracy were achieved at a small increase in cost. For example, the role-playing sessions that we have done required less than one hour.

Choosing the Method

This has been a long and complicated section. To help you in selecting a method for obtaining a judgmental forecast, Exhibit 6-7 lists the primary methods in the order that they were discussed, rating them according to criteria that are relevant to most forecasting problems. A high rating is favorable. Much subjective judgment was added to the evidence presented in arriving at the ratings. The ratings are relevant within a column; do not add across columns. The weighting for the criteria will depend upon the specific situation that is being forecast.

Of particular interest is whether the study involves intentions or opinions. For example, assume that the study involves an opinions forecast on a sensitive topic involving a large change, where there are conflicting parties, with little information about how others will behave. Assume further that this is an important situation involving large expenditures of money and that the forecast is not needed immediately. In such a case, sampling and nonresponse errors are not relevant, and one should have ample opportunity to ensure that the question is well defined. In this example, role playing would clearly dominate the other techniques. Incidentally, this is not a trivial case!

COMBINED FORECASTS

In judgmental forecasting, there is safety in numbers. The practice of obtaining forecasts from a number of judges and using the combination is a sound one. Research on this approach has been conducted since the early 1900s.

Certainly the value of combining forecasts was evident in intentions surveys. In this section, evidence is provided that it is good practice also for other techniques in judgmental forecasting such as opinions surveys, Delphi, and role playing.

The earliest studies on the value of using a combined forecast were
Exhibit 6-6  ROLE PLAYING VS. OPINIONS: ACTUAL SITUATIONS

<table>
<thead>
<tr>
<th>Situation</th>
<th>Conflict Among</th>
<th>Percentage of Correct Predictions (Sample Sizes)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Vietnam Bombing</td>
<td>Countries</td>
<td>Chance: 33, Opinion: 0 (1), Role Play: 100 (1)</td>
</tr>
<tr>
<td>USA-Mexico</td>
<td>Countries</td>
<td>Chance: 18, Opinion: 1 (1), Role Play: 57 (96)</td>
</tr>
<tr>
<td>Panalba</td>
<td>Stockholder &amp; Consumer</td>
<td>Chance: 20, Opinion: 41 (46), Role Play: 79 (57)</td>
</tr>
<tr>
<td>Philco Distribution</td>
<td>Manufacturer &amp; Retailer</td>
<td>Chance: 25, Opinion: 3 (33), Role Play: 80 (10)</td>
</tr>
<tr>
<td>Artists in Holland</td>
<td>Government &amp; Labor</td>
<td>Chance: 16, Opinion: 7 (14), Role Play: 50 (4)</td>
</tr>
<tr>
<td>NFL 55% Plan</td>
<td>Employees &amp; Owners</td>
<td>Chance: 33, Opinion: 27 (15), Role Play: 60 (10)</td>
</tr>
<tr>
<td>Unweighted Average</td>
<td></td>
<td>Chance: 25, Opinion: 21, Role Play: 70</td>
</tr>
</tbody>
</table>

*The sample sizes represent the number of predictions.
**Assuming that there are about three candidates available.
Exhibit 6-7  RATINGS OF METHODS TO OBTAIN JUDGMENTAL FORECASTS
(1 = unfavorable and 5 = favorable; n.a. = not applicable)

<table>
<thead>
<tr>
<th>Method</th>
<th>Cost</th>
<th>Speed</th>
<th>Situation Involves High Uncertainty &amp; Interaction Among Conflicting Parties</th>
<th>Questions Cannot be Well Defined</th>
<th>Sampling Error Is Important</th>
<th>Nonresponse Error Is Important</th>
<th>Judges Are Concerned About Evaluation of Their Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personal interview</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Telephone interview</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Mail questionnaire</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Delphi</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Traditional meeting</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>n.a.</td>
<td>1</td>
</tr>
<tr>
<td>Structured meeting</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>n.a.</td>
<td>3</td>
</tr>
<tr>
<td>Group depth interview</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>n.a.</td>
<td>2</td>
</tr>
<tr>
<td>Role playing</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>n.a.</td>
<td>5</td>
</tr>
</tbody>
</table>
conducted by asking people to estimate things like distance, volume, or weight. For example, Gordon (1924) had college students rank weights that they lifted; these rankings were then compared to the true order, and the average correlation for 200 judges was found to be .41. By averaging the rankings of any five random judges at a time, she obtained 40 combined rankings; the average correlation between the true ranking and the combined ranking was .68. Combined estimates were also obtained from groups of 10, 20, and 50 judges, and the correlations rose to .94 for the largest group. Similar results were obtained for judgments of such things as the temperature of a room, the number of items in a bottle, and the number of buckshot. DAWES (1977) demonstrated the predictive validity of combining in a problem calling for estimates of the height of people. Zajonc (1967) provides a summary of this literature.

It was concluded from these early studies that combining was at least as accurate as the average judge, almost always more accurate than the average, and sometimes better than the best judge. But Zajonc (1967) noted that these conclusions are not correct, technically speaking; they hold only if a majority of the individuals making judgments have a fair chance of being correct.

The possibility that a combination of forecasts can be better than the best forecast does not strike everyone as intuitively obvious. The demonstration is easy; one need only show that it is possible in at least one case. Consider that judge A predicts 10, judge B predicts 30, and the actual result turns out to be 20. The combined forecast of 20 in this case would be perfect even though each of the judges missed by 10.

Stroop (1932) duplicated Gordon’s (1924) results by having one individual make 50 judgments and taking the average of these judgments rather than using an average of 50 different judges. In other words, in certain situations, the average may be obtained by multiple judgments from a single individual.

Aggregation of one judge’s responses improves reliability, but not validity. When validity is a problem, it is better to average judges’ responses in the hope that the biases among judges will compensate for one another. Forecasts by different judges are expected to be especially useful in cases of high uncertainty. For example, Klugman (1945) found that combining led to improvements when trying to judge the number of unfamiliar items in a jar (lima beans), but produced no significant gain for familiar items (marbles). To carry this notion of uncertainty one step further, the use of a combination from different judges would seem especially relevant for long-range forecasting, where uncertainty is high.
The discussion to this point has assumed that the combination was based on an unweighted average. Winkler (1967a, 1971) examined different weighting schemes in his study of the prediction of football scores. He compared equal ratings, accuracy for the preceding week's forecasts, and self-ratings. He concluded that it did not matter much how the judgments were weighted. My conclusion from his data is that self-ratings and accuracy scores have modest value and can be used as a screening device to eliminate judges who have very low self-ratings or have demonstrated very poor previous accuracy relative to the rest of the group. This would presumably eliminate judges who either do not understand the question involved or do not possess a minimum level of expertise. An alternative that avoids issues of weighting and the dangers of outliers is to use the median. Medians would seem advantageous for comparisons across methods. However, SEWALL [1981] did not find medians to be useful.

The evidence on the value of combining is impressive, and it comes from a variety of areas. As long as the judges possess a minimum level of expertise, combining is better than the average accuracy of the component forecasts. Some of the early evidence is summarized here:

Thorndike (1938) asked 1200 subjects to make predictions about 30 future events. The subjects were asked to make an individual prediction and then a group prediction. The average individual was correct 61.9% of the time. The combination from groups of four to six people was correct 64.4% of the time. In cases where the majority was wrong, the combined prediction was worse than the average judge.

Klugman (1947) had 109 soldiers, during World War II, try to predict the dates for the ending of hostilities with Germany and with Japan. For the German armistice, the combination was better than 75% of the judges. For the Japanese armistice, it was better than only 46% of the judges. (In the Japanese situation, the judges were not aware of the atomic bomb, so most of them could be expected to be wrong.)

Kaplan, Skogstad, and Girshick (1950), in their study of forecasts of events in the social and natural sciences, found that the average
accuracy for all 26 of their judges was 53% with a range of 28% to 71%. Yet, if the choice favored by most forecasters had been used for each event, the accuracy would have been 68%, which was almost as good as the accuracy of the best judge.

Sanders (1963) used 12 students to make weather forecasts and found that the combined forecast was superior to the forecast by the best judge.

Goldberg (1965), in his study of psychiatric judgments, found that a combined staff score predicted almost as well as the best judge in a group of 29 clinicians.

Zarnowitz (1967) examined expert forecasts of the U.S. GNP and found that the group combination did better than 62% of the judges.

Winkler (1967), in forecasts of collegiate and professional football games, found that the accuracy of the combined forecast was almost as good as that of the best judge.


Gains are expected if one uses eclectic research; that is, in addition to using different judges, the other key factors in judgmental forecasting can also be varied. I was unable to obtain much evidence in this area. The two studies that were found provide only mild support because the sample sizes were small:
Levine (1960) presented forecasts of the annual U.S. investment in plant and equipment from 1949 to 1954, generated by two different surveys of intentions, one by the U.S. Department of Commerce and the Securities and Exchange Commission (SEC), and one by McGraw-Hill. I calculated a combined forecast by taking a simple average of the two forecasts. The Commerce-SEC survey had a mean absolute percentage error (MAPE) of 3.8; the McGraw-Hill survey had a MAPE of 4.3; and the combined forecast had a MAPE of 3.5. The differences were not statistically significant.

Okun (1960) presented forecasts derived from two intentions studies of the U.S. housing market. The Fortune survey of homebuilders and the Survey Research Center's (SRC) survey of buying plans were each used to forecast yearly housing starts from 1951 through 1956. A combined forecast was also developed from the average of these two forecasts. The SRC forecast had a MAPE of 8.5; the Fortune survey had a MAPE of 7.5; and the combined forecast had a MAPE of 6.5. The superiority of the combined method was not statistically significant.

ASSESSING UNCERTAINTY

One of the advantages of judgmental forecasting is that it provides a number of methods for assessing the uncertainty in forecasts. These methods are described here, and the empirical evidence on their effectiveness is examined.

The methods for assessing uncertainty can be grouped into two categories: those that ask the judge to assess the uncertainty of his forecast, and those that make comparisons among different forecasts. These methods can be applied to both intentions and opinions studies.

Self-Assessments

Please read the following sentence: "FINISHED FILES ARE THE RESULT OF YEARS OF SCIENTIFIC STUDY COMBINED WITH THE EXPERIENCE OF YEARS."

Now count the number of times the letter F appears in that sentence. Count them only once; do not go back and count them again. Record
your answer here: ______. Now state your confidence in your answer on a scale from 0%, meaning that you are sure you are incorrect, to 100%, meaning that you are sure you are correct. Record your confidence here: ______.

Most people feel confident of their answers. A convenience sample of 50 reported an average confidence level of about 91%. This confidence level proved to be unrelated to accuracy. For the 34% who had the correct answer (which is 6), the average confidence level was about 87%. For the 66% who had incorrect answers (which ranged from 2 to 5, with a mode of 3), the average confidence level was about 93%. Accuracy and confidence were unrelated in this study.

This letter-F test illustrates an important conclusion for self-assessment. Ratings of self-confidence by individuals are often of such poor validity that one should generally consider them to be worthless for predictions of a single event. This conclusion is difficult to accept personally, although it is a bit easier to accept where others are concerned. The evidence on this issue comes from a variety of areas:

Thorndike (1938) found a very small relationship between self-confidence and accuracy in his study using 1200 subjects to predict future events. The relationship was, however, statistically significant.

Kaplan, Skogstad, and Girshick (1950) found a rank correlation of only .2 between a measure of confidence provided by each judge and his accuracy in forecasting events in the social and natural sciences.

Holtzman and Sells (1954) found no relationship between self-ratings of confidence by 19 judges and accuracy in the prediction of the success of aviation cadets. This occurred despite high agreement among the raters.

Winkler (1971) found little relationship between self-ratings and accuracy in predictions of football scores by 55 judges.
In WARSHAW [1980] 60 housewives were asked about their intentions to purchase various brands of soft drinks over a five-day period. Global intentions were asked: “What is the probability that you will purchase brand X between now and next Monday morning?” Also, a decomposed version was asked of the same subjects with questions about location and about the purchase of multiple brands. The decomposed version (which Warshaw refers to as a “derived intention”) was significantly more accurate in predicting actual purchase behavior.

Prior research has shown that an eyewitness’s expressed confidence affects the extent to which people will believe the witness’s testimony. Jurors show much agreement in assessing the certainty expressed by the witness. It is interesting, then, to find that eyewitness confidence has little predictive validity. WELLS and MURRAY [1984] performed a meta-analysis on 31 previously published studies. The typical correlation between confidence and accuracy was only .07. This means that confidence explained less than 1% of the variance in accuracy. Wells says that some variables affect accuracy of judgment while other variables affect confidence. The results fit into a pattern of studies showing that individual confidence ratings are poor guides to accuracy unless the feedback in the task is exceptionally good. This pattern has a practical implication: don’t ask individuals how confident they are in their judgmental forecasts.

Although it is generally worthless to use the self-rated accuracy of individuals for single events, there is a small positive relationship. This means that self-rated accuracy should be useful for cases with many judges and, possibly, for a single judge making many predictions (LARSEN and REENAN [1979] and LINDLEY [1982]). Thus confidence ratings are useful for consumer intentions studies. Dalkey et al. provide evidence on self-ratings by groups:

Dalkey, Brown, and Cochran (1970), in a study involving almanac-type questions, found a strong relationship between average
Self-ratings of confidence for 16 groups (each with 15 to 20 judges) and the levels of accuracy for the groups.

Self-ratings are expected to be most useful when uncertainty is not great. Thus they would generally be more useful in estimating current status than in forecasting change. Evidence consistent with this viewpoint is provided by Best:

Best (1974), using Delphi, asked 28 faculty members at the University of Oregon's College of Business to estimate the current demand for the Oregon Business Review and also to estimate the current student enrollment in their college of business. There was a strong relationship between self-rated expertise and accuracy for each of the two estimates.

Self-confidence ratings are expected to be most useful in situations where the judge receives good feedback on predictive accuracy. This occurs in weather forecasting. The data by Cooke and by Williams showed high correlations between self-confidence and accuracy. The study by Ferber et al. was consistent, although the results were not statistically significant:

<table>
<thead>
<tr>
<th>Self-Rating of Forecast</th>
<th>Number of Forecasts</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Almost certain</td>
<td>685</td>
<td>98</td>
</tr>
<tr>
<td>Normal probability</td>
<td>970</td>
<td>94</td>
</tr>
<tr>
<td>Doubtful</td>
<td>296</td>
<td>79</td>
</tr>
</tbody>
</table>

Cooke (1906) rated the certainty of his weather forecasts in Australia in 1905:

Williams (1951) examined the records of eight professional weather bureau forecasters in making 12-hour forecasts of rain for Salt Lake City from November 1949 to January 1950.
Ferber, Chen, and Zuwaylif (1961) analyzed two-month manpower forecasts. The error of the forecasts was 4.7% for the 18 plants that kept records of accuracy and 5.3% for the 18 plants that did not keep records of accuracy.

Self-confidence is often poorly related to accuracy because "familiarity breeds confidence," but it does little to improve accuracy. Studies by Oskamp and by Ryback provide evidence in this area.

Oskamp (1965) provided biographical information about a person to 32 judges. The information was given in four stages. Stage 1 provided three sentences about the person's age, sex, marital status, race, military record, college background, and occupation; stage 2 provided one and a half pages on the person's life through age 12; stage 3 covered high school and college; stage 4 covered the army and other activities up to age 29. After each stage, the judge was asked 25 questions about the personality, attitudes, and behavior of the subject. No feedback was provided. The judge's accuracy was not significantly better than chance at any stage; however, their confidence went up markedly from stage 1 to stage 4.

Ryback (1967) asked 94 subjects to estimate the relative sizes of some geometrical figures. The judges repeated the task on five occasions spaced two or three days apart, with no feedback being provided to them. The accuracy of the judges did not change over time, but their confidence rose significantly ($p < .05$).
Another reason why self-ratings are poor is that they are **ipsative**. A low confidence rating for one person is not directly comparable to a low rating by another judge. The ratings are personal.

Estimates of a distribution are similar in many respects to ratings of self-confidence. A simple approach is to ask judges for “high,” “most likely,” and “low” estimates. The estimates might be defined as follows:

“What do you think would be the upper limit to the forecast? As the upper limit, choose the value which has less than 1 chance in 20 of being exceeded.”

A similar question would be used to secure a lower limit to the distribution. The upper and lower limits could then be viewed as the confidence interval.

Judges are overconfident, especially when they receive poor feedback as shown in the following studies:

Alpert and Raiffa (1968), in a study using almanac-type questions, found that the 98% confidence intervals stated by judges included the true value only about half the time. This overconfidence persisted even when the judges were given feedback about their overly narrow confidence bands, although the feedback was of some value. A replication of this study is provided in Pickhardt and Wallace (1974).

Staël von Holstein (1972) asked judges to forecast changes in prices of common stocks and to specify 99% confidence limits. Of all predictions, 43% fell outside these 99% confidence limits. Feedback was then provided to the judges that they were overestimating confidence. Of the predictions made after this feedback, 23% fell outside the 99% limits. Similar results were obtained by Schreuder and Klaassen (1984), but Fishhoff, Slovic, and Lichtenstein (1977) had less success with training.

One way to combat overconfidence is to ask forecasters to list explicitly the reasons why their forecast might prove incorrect [Cosier, 1978; Koriat, Lichtenstein, and Fishhoff 1980]. Another technique is to provide extensive training and personalized feedback.
Still another way to develop estimates of uncertainty is to use the principle of decomposition. Separate estimates can be made for each component of error: response, nonresponse, and sampling error. The total error may then be computed (e.g., by assuming that each type of error is independent of the other types and then adding variances). Brown (1969) presents a description of this approach, which he refers to as “credence analysis.” Although this approach has face validity, I have been unable to find any empirical studies that assess its value.

Comparisons among Forecasts

When judges make independent forecasts, the agreement among judges can be used as a measure of certainty. However, if the backgrounds are similar, as frequently happens in economic forecasts, agreement can be high while accuracy is low. The following two studies provide situations where agreement was not related to accuracy:

Holtzman and Sells (1954) showed high agreement among judges on which individuals would be successful as aviators after completion of a training program, but the accuracy of these forecasts was poor.

Ogburn (1934) found high agreement among football forecasts by sportswriters, but the accuracy was poor.

A striking example was provided by Baker (1957). Designers at the Ford Motor Company were asked to design the car that would best meet the needs of the consumer. The managers claimed that they were determined to be objective about everything. Baker reported as follows: “The final concept as it looked in plaster was satisfying to every designer in the company, and when you get 800 stylists under one roof to agree that they like a creation, you have unusually high agreement.” The company named this creation the Edsel. It was one of the biggest marketing disasters in the world.

Although agreement among judges can be a misleading measure of certainty, it is expected to be of value when the judges make truly
independent forecasts. In one study where the judges' forecasts were independent, a modest relationship was found between accuracy and agreement among judges:

Walker (1970) had subjects make estimates of the length of a line; the length, width, and height of a room; the weight of a book; the weight of a rock; the area of an irregular piece of paper; and the volume of a wastepaper bin. Four or more groups estimated each of the eight items, and each group had an average of 16 judges. Each judge worked independently. I reanalyzed Walker’s data to examine the relationship between the agreement within the group and the accuracy of the group. Agreement within each group was measured by calculating the coefficient of variation. Accuracy was measured by the absolute percentage difference between the average group estimate and the true value. When the coefficient of variation for the group’s estimate was less than 10%, the mean absolute percentage error (MAPE) was 7%. When the coefficient of variation exceeded 10%, the MAPE was 19%.

In cases where one cannot make comparisons among judges, repeated forecasts can be obtained from the same judge. The measures should be separated in time, generally by days or weeks, and an examination should be made of the consistency over time. Such an approach is useful when the number of judges is restricted by a need for secrecy.

The use of eclectic research should, once again, be considered. Uncertainty could be assessed by comparing forecasts obtained by different judgmental methods. “Different methods” means using different forecasters, different phrasing of the question, different ways of collecting the forecasts, or, preferably, all of these. The basic assumption here is that each of the different approaches is subject to different errors. Disagreement among the results obtained by the various approaches should lead to less confidence in the forecast. This makes sense but it is pure speculation; I found little empirical evidence on this issue.

The various methods for assessing uncertainty are listed in Exhibit 6-8, along with guidelines for their use and some guesses about their costs. It is advisable to use more than one method for assessing uncertainty. Remember: ECLECTIC RESEARCH ALWAYS.
### Exhibit 6-8 RATINGS OF JUDGMENTAL METHODS TO ASSESS UNCERTAINTY

<table>
<thead>
<tr>
<th>Method</th>
<th>Uncertainty Is</th>
<th>Feedback Is</th>
<th>Number of Judges Is</th>
<th>Cost Is</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Self-assessments</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-confidence</td>
<td>Low</td>
<td>Good</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>Distribution estimates</td>
<td>Low</td>
<td>Good</td>
<td>Large</td>
<td>Low</td>
</tr>
<tr>
<td>Credence analysis</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Moderate</td>
</tr>
<tr>
<td><strong>Comparisons among forecasts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Different judges</td>
<td>High</td>
<td>Poor</td>
<td>Large</td>
<td>Moderate</td>
</tr>
<tr>
<td>Time consistency</td>
<td>High</td>
<td>Good</td>
<td>Small</td>
<td>Moderate</td>
</tr>
<tr>
<td>Different methods</td>
<td>High</td>
<td>Poor</td>
<td>Small</td>
<td>High</td>
</tr>
</tbody>
</table>
SUMMARY

Intentions and opinions data provide the two basic types of information used in judgmental forecasting. Intentions forecasts are relevant for events that are important to the judge. Assumptions are made that a response can be obtained from the judge, that the judge actually has a plan, that he reports correctly, that he can fulfill the plan, and that he is unlikely to change the plan over the forecast horizon. These assumptions imply that intentions data are most useful for assessing current status, somewhat useful for short-range forecasting, and of little value for long-range forecasting. Opinions data are simpler to obtain and can be used in a wider variety of situations.

Numerous errors arise in judgmental forecasting. One of the most serious is bias. Judges confuse their desires for the future with their forecasts. Another major problem is anchoring; judges are conservative and bound by tradition.

The development of judgmental methods was described in three stages: selecting judges, posing the question, and obtaining the forecast. Some practical suggestions were provided in each of these areas to eliminate problems of bias and, to some extent, to compensate for anchoring. The more important suggestions are summarized here.

On Selecting Judges

1. For intentions studies, use probability samples.
2. For opinions studies:
   (a) Use inexpensive experts to forecast change (assuming that they have at least a minimum level of expertise in the area to be forecast).
   (b) Do not use judges who are personally involved in the situation being forecast.

On Posing the Question

1. Use decomposition when:
   (a) Uncertainty is high,
   (b) Prior theory exists,
   (c) Different judges have different information.
2. Word the question in different ways, especially in cases where uncertainty is high (as in long-range forecasting).
3. Provide only the minimum relevant information to the judge.
4. Organize the information so the judge can understand it easily.
5. Use scales with many gradations for the judge’s responses. Five gradations are O.K., but 11 are better. Use an odd number of scale points.
6. Ensure that probabilities of totally exhaustive and mutually exclusive events sum to 1.0
   (a) Use two way splits in the questions
   (b) Emphasize to the respondent that the probabilities should sum to 1.0
   (c) Normalize when necessary
7. Consider projective questions for sensitive issues.

On Obtaining the Forecast

1. In selecting a method to obtain forecasts, consider both “judge-judge” and “researcher-judge” interaction. Exhibit 6-5, “Interaction Effects on Techniques to Obtain Judgmental Forecasts,” is provided as a guide for selecting appropriate methods.
2. A checklist for the selection of the most appropriate method to obtain forecasts is provided in Exhibit 6-7, which rates eight techniques against seven criteria.
3. The mail survey provides an inexpensive way to obtain intentions forecasts, and it is useful for sensitive issues.
4. Traditional group meetings are the most popular method of obtaining important judgmental forecasts, apparently because group satisfaction is higher with unstructured group meetings. Unfortunately, this is the poorest way of obtaining judgmental forecasts.
5. Structured group meetings are more effective than traditional meetings and cost about the same.
6. Role playing is useful when there are conflicting groups and uncertainty is high. It provides more accurate long-range forecasts for important events.

On Using Combined Forecasts

1. Use combined forecasts! The combined forecast is almost always better than the average component from which it was derived, and it is sometimes better than the best component.
2. Combining is especially valuable when uncertainty is high (as it is in long-range forecasting).
3. Use eclectic research; base the combination upon judges, questions, and methods of obtaining forecasts that differ substantially.

On Assessing Uncertainty

1. The basic approaches to the use of judgmental methods in assessing uncertainty are to ask the judges to rate their confidence, and to make comparisons among different judgmental forecasts.
2. Judges tend to be overconfident. This is especially true when experience is high and feedback is poor.
3. Self-ratings of confidence are valuable in situations where the judges have been receiving good feedback on the accuracy of their forecasts, and where large groups of judges are used.

In general, common sense in the use of judgmental methods of forecasting has led to undesirable practices. The most common approach is to use biased judges and have them make forecasts in traditional group meetings. A number of simple changes can lead to great improvement in accuracy at a modest cost. We think it is ridiculous when others use the crystal ball, but when we gaze into its depths . . .