LONG-RANGE FORECASTING
From Crystal Ball to Computer
### Part IV

**COMPARING METHODS**

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Costs and Benefits of the Forecasting Methods</td>
</tr>
<tr>
<td>15</td>
<td>An Evaluation of Accuracy</td>
</tr>
</tbody>
</table>

Part IV compares the forecasting methods and answers the question, "Which is the best method for a given situation?" Chapter 14 examines the costs and benefits of the various methods in general terms. Chapter 15 evaluates the relative accuracy of each method.
LONG-RANGE FORECASTING
From Crystal Ball to Computer
Fourteen

COSTS AND BENEFITS
OF THE FORECASTING
METHODS

Contents

Developmental Costs .......... 367
Maintenance Costs .......... 370
Operating Costs .......... 371
Assessing Uncertainty .......... 372
Assessing Alternative Futures .......... 372
Learning .......... 374
Problems with Judgmental Learning .......... 375
Improving Judgmental Learning .......... 379
Summary .......... 385
Men occasionally stumble over the truth, but most of them pick themselves up and hurry off as if nothing had happened.

Winston Churchill
(Noted in Cetron and Ralph, 1983)

This chapter presents a checklist to help in the selection of a forecasting method. The assumption is that an explicit evaluation of costs and benefits is superior to a subjective global assessment. The latter approach is likely to lead to the conclusion that the new forecasting problem looks just like the last one. Forecasters, like other people, get into ruts. An explicit evaluation gives other methods due consideration.

Following the cost-benefit framework of Chapter 11, I rated the various forecasting methods. Because it is so important, accuracy is treated separately in the next chapter. My ratings were made at two different points in time; on average, each of these ratings was within 0.5 rating points on the five-point scale. In addition, Robert Fildes, who has published comprehensive evaluations of forecasting methods [e.g. FILDES, 1982], provided independent ratings. His ratings and my most recent ratings were within 1.0 rating points on average. The ratings in Exhibit 14-1 are based on an average of my two sets of ratings and the ratings by Fildes.

Ratings of forecasting methods also depend upon the particular forecasting problem. For example, what data are available? What capabilities exist in the organization for using various methods? The ratings provided in Exhibit 14-1 do not consider differences in the situation. You might want to revise the ratings to suit your particular problem. You could even start from scratch and enter your own ratings, so that you avoid problems with anchoring.

Once ratings are obtained, it is necessary to decide how to weigh the costs and benefits. This task is not difficult on the cost side; however, the assessment of benefits usually presents problems. Is it more important to obtain a good estimate of uncertainty or to be able to assess alternative futures? Is it more important to be able to assess alternative futures or to improve accuracy? These questions are specific to the situation, so I leave the difficult work to the reader. Here is an example of the type of analysis:
Selecting Forecasting Models: An Example

Assume that methods were being considered to provide long-range forecasts of automobile sales for General Motors. Ideally, one should treat each forecasting method as an investment and calculate a rate of return. The various column headings from Exhibit 14-1, along with accuracy, would enter into these calculations. If difficulties were found in translating the estimates into a common unit of measure (say dollars), one could use a satisficing model, that is, minimum acceptable levels would be used for each criterion. Comparisons among models could also be made by rating the importance of each criterion. For example, a 1 to 5 importance rating (5 being most important) for the General Motors example might result in the following: 1 for the cost factors, 3 for assessing uncertainty, 5 for assessing alternative futures, and 4 for learning. These importance ratings would then be multiplied by the ratings in Exhibit 14-1. The benefits due to accuracy could be assessed by using the framework in Appendix A.

The rest of this chapter covers the major considerations used in rating costs and benefits. The discussion covers each of the criteria listed in Exhibit 14-1.

DEVELOPMENTAL COSTS

Three major factors influence developmental costs:

- Data needs. Methods that require more data cost more because of added collection and processing expenses.
- The complexity of the method. More complex methods require more highly trained people and more time for analysis.
- Implementation. For both analyst and user to gain confidence in the model, significant time and money must be invested.

Thus complex methods have much higher development costs.

Data needs, complexity, and implementation were each examined to obtain rankings of development costs for the methods listed in Exhibit 14-1. As may be seen, judgmental methods are the least expensive because they draw mostly upon existing capabilities and existing data. Objective methods often seem deceptively simple. When rating developmental costs or when preparing the forecasting budget using objective methods, remember Armstrong’s law: “There’s no such thing as an easy job.”
<table>
<thead>
<tr>
<th>Method</th>
<th>Costs</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Development</td>
<td>Maintenance</td>
</tr>
<tr>
<td><strong>Judgmental</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intentions</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Opinions</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Delphi</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Traditional meeting</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Structured meeting</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Group depth interview</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Role playing</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td><strong>Extrapolation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Markov chains</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Moving averages</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Exponential</td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoothing</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Box-Jenkins</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Regression</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>A priori</td>
<td>Updated</td>
</tr>
<tr>
<td>---------------</td>
<td>----------</td>
<td>---------</td>
</tr>
<tr>
<td><em>Econometric</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A priori</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Updated</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Segmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trees</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Input-output</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Simulation</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Segmentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trees</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Input-output</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Simulation</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Bootstrapping</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Direct</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Indirect</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

*The ratings may be compared only within columns (e.g., a 4 on uncertainty is not related to a 4 on learning).*
Remember the advice presented throughout Part II: keep it simple. Things will become complex all by themselves. Of course, we academics always like to make things look difficult. Perhaps that is our way of getting respect . . . and money. Out West they build dams to tap the pork barrel. The South constructs the nation’s military defenses. Here in the East we do complex research for the government. One of my more memorable experiences was the difficulty in getting permission to use a simple method when a complex one could also be used; the job involved the federally funded Minicar project, and it eventually led to the paper by Armstrong and Overton (1971). (The hassle reminded me of the three men who went to lunch in Washington, D.C. The bill came to $100. The first man reached for the check, but the second man snatched it away from him saying, “I’ll put it on my company’s expense account because it will only cost $50 after the tax deduction.” Ultimately, however, the third man took possession of the bill; his firm was on a cost-plus government contract so that it made $20 on the lunch.) But I digress. The point is that complex methods are often encouraged simply because they are expensive and impressive. Avoid this.

MAINTENANCE COSTS

The argument for maintenance costs parallels that used for developmental costs. Complex methods are more expensive to maintain. They require more data and more effort to document and keep up to date. Thus the ratings are similar to those for developmental costs.

Often data are routinely collected for other purposes, so the full cost should not be charged to forecasting. Still, expenses are involved to ensure that the data are collected in a consistent manner. Updating and revising the data create small additional costs. A major concern is changing definitions. These may call for expensive revisions. Unfortunately, shifting definitions are common: “Gee, we changed the definition of a riot in 1967,” “Alaska and Hawaii were added to our totals as of 1968,” “Foreign cars were added to the totals starting in 1958,” and so on.

For maintenance costs, remember Murphy’s law: “If something can go wrong, it will.” In deference to Murphy’s law, you should keep the methods simple.
Unlike developmental and maintenance costs, operating costs usually are not much higher for complex methods. Once a complex objective method is developed, operating costs are low because the forecasts are generated either by computer or by a clerk performing routine calculations. (Of course, some consultants do not like the client to get away that easily, so they may develop objective methods that require continuing assistance.

Operating costs are usually high for judgmental methods. Judgmental forecasts have short life spans and before long, new forecasts must be obtained. The forecasts generally call for time expenditures by analysts, managers, and other well-paid people.

Operating costs are of particular interest when a large number of forecasts are required. For example, in inventory control problems, forecasts are often required for thousands of items. Operating costs are also important if forecasts are desired for alternative assumptions about the future environment or for the examination of different policies.

Low operating costs are preferable for pragmatic reasons. Organizations view forecasting as a staff activity that is not crucial for day-to-day operations. In bad times, organizations save money by cutting frills... such as forecasting. For example, I once had a consulting job with a large firm whose sales had taken a sudden turn for the worse. I worked with a man named Jones to develop better short-range forecasting methods. On Monday, Jones and I worked to prepare for a presentation on Friday. On Wednesday, I called Jones to see whether everything was ready to go. The operator answered, “I’m sorry; we have no Mr. Jones here.” Jones, it seemed, was fired on Tuesday. It is not clear that Jones deserved this, and he certainly had not predicted his own fate. But Jones was the only forecaster in an organization that seemed to be in dire need of better forecasts. (Come to think of it, this reflects poorly on my accuracy as a forecaster because I did not predict that Jones would be fired. But, then, I too suffer from optimism and anchoring.)

Parkinson’s law (Parkinson, 1957) is worth remembering for operating costs. It says that “work expands so as to fill the time available for its completion.” A corollary would be that “forecasting expenses rise to meet the budget.” This is especially true for judgmental methods. According to Morgenstern, Knorr, and Heiss (1973), some organizations have spent over $250,000 for a Delphi study.

I have looked, without success, for evidence that judges make more
accurate forecasts if they spend more time on each forecast. Negative findings were obtained in Hall, Mouton, and Blake (1963) and in Chapman and Chapman (1967). This lack of evidence does not bother judges. They still act as if they can do a better job, given more time. As an extreme example, Schneidman (1971) took four months to predict which of 25 subjects was most likely to commit suicide. Gough (1962), in his review of the literature, found most writers agreeing that the judgmental method is a time-consuming, painstaking task; he refers to a study by Sanford, who claimed that a minimum of six hours is required to predict the success of a student entering college. Note that judgmental methods lead to high operating expenditures.

**ASSESSING UNCERTAINTY**

The use of each method for the assessment of uncertainty was discussed at length in Part II of LRF. Here are some generalizations:

- Judgmental methods are advantageous because of the different ways one can estimate uncertainty, but they tend to understate uncertainty.
- Extrapolation methods offer simple and inexpensive ways to assess uncertainty.
- Econometrics and segmentation can, in some cases, go beyond extrapolations to explain the sources of uncertainty.

More emphasis should be given to the assessment of uncertainty. RUSH and PAGE [1979] found that the use of measures of uncertainty in published forecasts declined from 1910 to 1964. Firms surveyed in DALRYPLE’s [1985] study typically reported little or no formal assessment of uncertainty.

**ASSESSING ALTERNATIVE FUTURES**

The analysis of alternative futures, or sensitivity analysis⁶, is useful for planning purposes. Questions such as “If X occurred, how would people respond?” can be asked. The organization can then try to influence whether or not X occurs; or, if they have no control over X, they can plan for this contingency. This section discusses the question of which forecasting method is best for sensitivity testing.

Extrapolation is of little value for examining alternative possibili-

---

⁶ Sensitivity analysis is a method for predicting how changes in the input variables of a model will affect the output variables.
ties. On the other hand, segmentation and econometric methods are well designed for this; the inputs can be varied to reflect the changes in the environment or the changes in the organization's policies, and the results can then be examined.

Econometric methods are particularly well suited to testing alternative futures because the alternatives can be examined in a consistent and inexpensive manner. One merely inserts different values for the causal variables to reflect the changes in the environment or in organizational policies. It is simple to vary the inputs, and the outputs are easy to interpret. An example of this use of the econometric method is provided in my study on the photographic market:

In Armstrong (1968a), an econometric model was used to assess the impact of changes that were expected to influence the photographic market. An analysis of the Kennedy-round tariff cuts showed that sales would increase substantially in underdeveloped countries (e.g., a 37% increase for Brazil), but only modestly in developed countries. Other analyses could also be made: What would happen if resale price maintenance were abolished in Austria? What if the high sales taxes in Greece were reduced? What if economic growth in the United States slowed down substantially?

Judgmental methods can be used to assess alternative futures, but the estimates may be adversely affected by bias and anchoring. Furthermore, a group may become incensed at the introduction of alternative possibilities because its members see this as an attack on their good judgment. As a result, it helps to use unbiased judges (people who are outside the influence of the organization). Efforts should be made to reduce the influence of the researcher by using structured methods.

Groups can adopt formal procedures to remove the detrimental effects of their norms (assuming they can decide that there are detrimental effects). Two techniques are of particular value:

1. Group depth interviews are useful for predicting how members of an organization will act in a given situation. For example, how would they act if their factory were closed?
2. Role playing goes beyond group depth interviews in the assessment of alternative futures by protecting the participant from peer pressure (he is only acting within his new role!). New situations can be
Costs and Benefits of the Forecasting Methods described, including new types of roles, new organizational structures, new decisions, or new environments. The consequences of these new situations can be acted out. For example, role playing has been used to examine reactions to new offers or new strategies in labor-management negotiations, to examine the jury's reaction to a defense strategy in a law case, and to test customers' reactions to a new marketing strategy.

LEARNING

You were probably wondering whether I had forgotten about Winston Churchill and people who stumble over the truth. Not at all. This section discusses truth and learning. I am a firm believer in progress; a forecasting model should be improved over time.

The extent to which forecasting models contribute to learning varies greatly. Extrapolation methods can be quickly discounted. The researcher may learn from experimentation which method works best, but the reasons for the improvement seldom are obvious. If the environment changes, it may be necessary to develop a new extrapolation model.

At the other extreme, the econometric model is well suited to learning. Analyses of conditional forecasts are useful in indicating where learning is needed. Additional information can be used to update the relationships in the model.

The mere presence of the econometric model serves as a focus for learning. Existing knowledge is conveniently summarized by the model. Furthermore, the learning is no longer dependent upon key personnel; they may leave the organization, but the model will remain faithful. As described in the following example, econometric models can indicate areas where we need to learn more and they provide a focus for this learning.

In my study of the photographic market (Armstrong, 1968a), the development of the econometric model revealed great uncertainty about price elasticity. Furthermore, there was a lack of data about camera prices in various countries. Communication with top management in some large photographic companies led to the conclusion that such data were not obtained in a systematic way. Some studies had been done on price elasticity, but because they
lacked the focus of a formal model, they had been filed away. Management forgot about the results.

Segmentation methods also offer a framework for learning. Survey results can be added to these models, rather than having the results filed in the warehouse or in the waste basket.

Judgmental methods are of particular importance because they are used for most important forecasts, and because we know a lot about the learning process that occurs in judgmental methods. In the rest of this chapter, I examine why judges have difficulty learning and what can be done to improve learning.

**Problems with Judgmental Learning**

Expertise . . . breeds an inability to accept new views.

Laski (1930)

Consider the following study. Skinner (1948) put a hungry pigeon in a cage. On a certain time schedule, food was supplied to the pigeon. This time schedule was fixed and had nothing to do with the bird. What happened? The bird “learned” how to make the food appear. Whenever it wanted food, it repeated the behavior that it was engaged in when the food first appeared. For example, if the pigeon was turning counterclockwise when the food appeared, it concluded that its counterclockwise movement had produced the food. This initial learning proved to be highly resistant to change, even though it had nothing to do with the appearance of the food.

Do people do any better than pigeons? Consider the stockbroker. Consider also the manager—Strickland did, and people did a good job of simulating pigeons. Kahneman and Tversky imply that we act like pigeons.

Strickland (1958) had subjects act as managers for two subordinates whom we will call Stan and Ned. The manager could see Stan’s work and could communicate easily with him. Communication was poor with Ned. Over the total work period, both Stan and Ned produced the same amount of work. Whom did the man-
ager trust? He trusted Ned; Stan, he thought, required constant surveillance to produce the same output. In other words, the manager concluded that his managing was largely responsible for Stan's output. (This result is also interesting because increased communication led to less trust.)

Kahneman and Tversky (1973) discussed a training program for a flight school. Trainers adopted the recommendation from psychologists that they use only positive reinforcement for training; that is, they praised successful work and said nothing otherwise. After a time they concluded that positive reinforcement did not work; they would praise someone for successfully completing a series of complex maneuvers, but this trainee would not do as well on the next trial. What was happening? Learning involves making some mistakes. The student cannot perform successfully on each trial. Thus you expect regression toward the mean; an exceptionally good trial will be followed by a more average trial, and similarly for an exceptionally poor trial. The flight school trainers saw this regression toward the mean, but they attributed these changes to their actions as trainers. As a result, they learned that "what works" was to punish someone for bad behavior because then he would probably improve on the next trial. Rewarding others, they concluded, just led to overconfidence on the part of the learner.

The illusion of control occurs even in situations where the person clearly has no control, such as in gambling (Langer, 1975). Mark Twain said it well in describing a fight: "Thrusting my nose firmly between his teeth, I threw him heavily to the ground on top of me."

In Chapter 6, evidence was presented to show that experts do not learn from experience. The pigeon-type studies provide a clue as to why this occurs: people and pigeons sometimes use a poor strategy for learning. The major problem is that they look for confirming evidence rather than disconfirming evidence. A bit of confirming evidence may help to reach a minimum level of expertise. After that, it does not help. Wason's studies provide evidence on this issue:
Wason (1960, 1968a) presented subjects with a three-number sequence: 2, 4, 6. The subjects were told that this sequence had been generated by a rule that the experimenter had in his head. The subjects were then asked to learn the rule by generating additional three-number sequences (e.g., 8, 10, 12). After each sequence, the experimenter told the subject whether or not the new sequence agreed with the rule. The subject could generate as many three-number sequences as he wished; when he felt confident of the rule, he wrote it.

The correct rule in the Wason study was “three numbers in increasing order of magnitude,” that is, \( a < b < c \). Only about 25% of the subjects learned the correct rule. (I replicated this experiment a number of times and my results have been similar to Wason’s.) Usually a subject selects a hypothesis (e.g., “add two to each successive number”) and looks only for evidence to confirm this hypothesis (e.g., 10, 12, 14). He does not attempt to refute his hypothesis. In other words, most people refuse to entertain the possibility that they are wrong!

The story gets worse. Subjects who wrote the wrong rule were allowed to try again (i.e., they were allowed to generate additional sets of numbers to obtain more evidence). About half of these subjects continued to search for confirming evidence for the same rule. (It’s like magic; if only we pronounce the rule correctly, then it will work.)

It is not clear whether subjects failed to use disconfirming evidence because they were unable or because they were unwilling. When asked how to find out whether their hypothesis was wrong, however, few of them recognized the need to look for disconfirming evidence by generating a sequence of numbers inconsistent with their hypothesis.

Before you try this test on others, remember that it can be threatening to people’s self-esteem. So do not ask them to reveal their answer. One of Wason’s subjects reacted in an extreme fashion and had to be removed by ambulance.

The above studies are of immense importance to the issues discussed in \textit{LRF}. They imply that:

1. Rational arguments will not be successful in implementing important changes (remember also the Jesus Christ study from \textit{LRF} p. 24).
2. Expertise offers little advantage in forecasting the effects of large changes.
3. Anchoring and bias will probably occur in judgmental forecasting.

So . . . are the studies correct? Now I am in a difficult situation. If you believe “yes,” I can present confirming evidence to make you happy. If you believe “no,” I can use the rational argument and you may get mad at me. So I will not say much. Of course, you could learn more on this from the substantial amount of literature now available.* The following study by the Chapmans also suggests that people seek confirming evidence:

Chapman and Chapman (1969) asked 32 experts to examine data from homosexual and heterosexual subjects. The data on these subjects were contrived so that there were no relationships for variables that previous empirical literature had found to be irrelevant. Nevertheless, the practicing clinicians saw the relationships that they expected to see which, incidentally, were the same invalid relationships that were expected by a group of nonexperts. Furthermore, some valid relationships conflicted with folklore in this case; when these valid relationships were introduced into the data, the clinicians still saw the invalid relationships. In other words, they had great difficulty in seeing the valid relationships in the data even though their effects were large. (Similar results for a different problem were obtained by Chapman and Chapman, 1967.)

The Chapmans’ study suggests that experts do not learn effectively when disconfirming evidence is given to them. Strickler also found that experts did not learn from disconfirming evidence, but nonexperts did:

*Further evidence is also provided in Bruner and Potter (1964), Wason (1968b, 1969), Johnson-Laird and Wason (1970), Hartsough (1975), Langer (1975), Langer and Roth (1975), MAHONEY and KIMPER [1976], MAHONEY and DeMONBREUN [1977], MYNATT, DOHERTY and TWENEY [1978], LORD, ROSS and LEPPER [1979], MANTTELOW and EVANS [1979], JONES and RUSSELL [1980], and TWENEY and YACHANIN [1984].
Strickler (1967) replicated a study by Hiler and Nesvig (1965). The task was to predict which figure drawings had been done by psychiatric patients and which by normal people. Hiler and Nesvig found no difference between eight students and six clinicians in their ability to predict, as the former were correct 65% of the time, and the latter were correct 64% of the time. Strickler added a new wrinkle; before the judges were asked to make their predictions, they were given results from an empirical study on aspects of drawings that were related to psychiatric problems. Six practicing clinicians, all with Ph.D. degrees and with an average of 14 years of experience, were correct on 66% of the 87 drawings. In other words, they did not do any better with this new information than they had done in the Hiler–Nesvig study. However, the students were correct on 72% of the cases, which was significantly better than the clinicians’ level of accuracy (p < .05). The inexperienced judges did learn from the new information.

Improving Judgmental Learning

Bernard Baruch was asked, “What is the secret of your success at being economic advisor to seven Presidents of the United States? He replied, “It’s rather simple. People ask my advice because I have good judgment. Good judgment comes from experience. Experience—well, that comes from bad judgment.”

(Noted in Firestone, 1972)

The preceding section reviewed problems with judgmental learning. These were caused by factors relevant to the situation as well as to the judges. An examination of these factors will help identify areas for improvement. The factors include the following:

<table>
<thead>
<tr>
<th>Situation</th>
<th>Judge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>Feeling of expertise</td>
</tr>
<tr>
<td>Feedback</td>
<td>Search for disconfirming evidence</td>
</tr>
</tbody>
</table>

Each of these factors is examined in the following discussion. For a more complete discussion, see FISCHHOFF and MacGREGOR [1982].

Complexity. Evidence from Archer, Bourne, and Brown (1955) and from Dudycha and Naylor (1966) suggests that learning proceeds more
slowly when the task is complex. Although you can seldom change the task, its apparent complexity can be reduced by using decomposition. Decomposition can explicitly organize the bits and pieces of subjective information so that the judge can use feedback more effectively. The mere use of paper and pencil is one step in this direction. For example, in the study cited by Chapman and Chapman (1967), the accuracy of the judges was improved by providing them with paper, pencil, and a ruler. Smedslund (1963) also found a slight gain from paper and pencil, although most subjects did not even think to construct $2 \times 2$ tables for their data, which consisted of two variables, each variable having two possibilities.

**Feedback.** Ryback (1967) reviewed the literature on the value of feedback and found that when feedback was absent, little learning occurred. Conversely, he found that the more immediate and precise the feedback, the greater the learning. Not surprising, MURPHY and DAAN [1984], found that feedback led to improved accuracy by weather forecasters.

Feedback is most useful when it is tied to the individual's own behavior, with negative feedback for incorrect responses, and positive feedback for correct responses. This is much more effective than general advice, as shown by GAETH and SHANTEAU [1984].

Although feedback should be rapid, it is also important that it be summarized in a way that is easy to understand. Case-by-case feedback was helpful in Sechrest, Gallimore, and Hersch (1967), but it was not effective in SCHMITT [1978] or FISCHER [1982] nor in the following study:

In Graham (1971), case-by-case feedback was provided to novice judges as they used personality data (from the MMPI) to predict which persons were college students and which ones were hospitalized psychiatric patients. The 14 judges who received case-by-case feedback were correct on 72% of the cases. This was better than chance (50%), but it was not significantly better than the results for the 14 judges who received no feedback (69%).

Data that are summarized after a number of trials are more effective. Hammond, Summers, and Deane (1973) found no gain for case-by-case feedback, but feedback after each group of 20 trials improved results. The following study provides further support:
Ward and Jenkins (1965) presented contrived data on the relationship between cloud seeding and rain to a group of 72 judges. The data can be summarized as follows:

<table>
<thead>
<tr>
<th></th>
<th>Rain</th>
<th>No Rain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seeding</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>No Seeding</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Previous studies cited by Ward and Jenkins suggested that, when data are presented on a case-by-case basis, the judgment of a relationship relies primarily on the frequency of entries labeled $a$ above, that is, cases where the policy variable (seeding) and the intended outcome (rain) occur together. Thus seeding in a wet climate would lead a judge to perceive a relationship even though seeding may have no relationship to rain. In other words, the judgments are not based on conditional probabilities, as they should be. In this particular experiment, it was found that only 17% of the judges followed a logically defensible rule in perceiving the data when receiving feedback on a case-by-case basis; however, when summary information of the type outlined in the seeding-rain table was provided, 75% did so. In fact, subjects receiving only summary data did better than those receiving both case-by-case and summary data.

The preceding discussion on feedback considered only information on the accuracy of predictions. You might also consider other types of feedback. The major types are illustrated by the Brunswick lens model of Exhibit 14-2. This model, taken from Brunswick (1955), has been adapted to fit the terminology used in LRF. The $b$'s represent the estimated relationships according to the actual data, while the $\hat{b}$'s represent the relationships as seen by the judge. The dashed line represents feedback on the accuracy of the judge’s predictions. (WERNER, ROSE, and YESAVAGE [1983] demonstrate the Brunswick Lens Model in forecasting dangerous behavior by mental patients.)

The bootstrapping model can provide feedback to the judge on how he is making the decisions. The econometric model provides information on the actual relationships. Actual outcomes and a record of forecasts are needed to assess accuracy.

What type of feedback is most useful? According to Newton’s study, feedback from the econometric model improves accuracy the most. This
Costs and Benefits of the Forecasting Methods

Exhibit 14-2 THE BRUNSWICK LENS MODEL OF FEEDBACK

Notes:
1. The X's are the causal variables.
2. The dashed line represents feedback on the accuracy of the judge's predictions.
3. The solid lines represent relationships.

adds support to Goldberg’s (1959) speculation, drawn from a sample of one. (Support is also provided in Nystedt and Magnusson, 1973):

Newton (1965) had judges make predictions of 53 students’ grade-point averages from information on IQ, high school rank, college board score, and a rating by the high school principal. Five feedback conditions were established. In each case, information was provided on accuracy. Cases 2 through 5 also received the feedback from the following sources; thus, the conditions were:

1. Accuracy only
2. Bootstrapping model and accuracy
3. Econometric model and accuracy
4. Bootstrapping and econometric models and accuracy
5. Same as case 4 but with more explicit rules
The feedback from the econometric model (condition 3) yielded significant though small improvements in accuracy. This seemed to be the only important factor in this study. Feedback conditions 1 and 2 yielded no improvements in accuracy, while conditions 3, 4, and 5 each provided similar improvement.

Feelings of Expertise. Evidence presented in Chapter 6 suggested that education and experience generally have a detrimental influence on forecasters because they increase confidence but not accuracy. This increased confidence interferes with further learning; this, in turn, may reduce accuracy, although evidence here is inconclusive (e.g., Crow, 1957).

It does not seem like a friendly act to tear down a forecaster's confidence. However, you can save money by avoiding expensive educational programs. The section on feedback concluded that the key element in learning about the forecast situation is empirical evidence. Thus, if there is little empirical evidence, avoid training in that area; all you will do is transmit the folklore. The discussions on the virtues of simplicity also indicated that there is no need for expensive training programs. You can simply use empirical studies. Now, how many highly trained real-world forecasters do you know who use this approach to learning? As nearly as I can tell, it is a rare few. Instead, expensive programs that transmit folklore are preferred. (If the same standards were applied to university professors as are applied to the drug industry, we would be out of our jobs tomorrow. We have tried, but have been unable to prove efficacy. We can't prove that we help people to learn more effectively than they could learn by living a more normal life. But that's another story.)

Remember Lem Putt, the privy builder (from Chapter 6)? Expertise is sometimes useful and education in some areas is relevant. In addition to privy building, worthwhile areas include how to assess the current situation and how to improve certain skills (e.g., skills in using forecasting methods).

Search for Disconfirming Evidence. An active search for disconfirming evidence can greatly improve learning ability—even for us experts. Disconfirming evidence becomes less threatening when we are in control, and when we seek it out. What happens when active search is absent? The Jesus Christ study (Batson, 1975) suggested that disconfirming evidence increased the experts' confidence in their currently held beliefs. This finding was supported also by Geller and Pitz (1968).
One of the problems is that judges remember incorrectly, as illustrated in the following study:

Fischhoff and Beyth (1975) asked judges to make predictions on political and social events. After the events had passed, they went back to the judges and asked them what predictions they had made. The judges were seldom surprised by the outcomes. Often they remembered incorrectly what they had predicted. Even when they had written their predictions, and the predictions could be seen to be incorrect, they rationalized what was written and claimed, “I knew it would happen.” This type of remembering is convenient for us; it means that we can go through life without changing our beliefs. (More on this study is provided by Fischhoff, 1975, and hindsight bias by physicians is shown in ARKES [1981].)

Unfortunately, learning is painful. Finding one’s mistakes is not a pleasant experience. (Do you think I like having to add page 450 of this edition, the one that lists errors in the first edition?) In groups, learning gets even more painful. As a result, many organizations have procedures to discourage learning. They focus on finding evidence to confirm all of their prior decisions and actions. Many organizations and governments have achieved the level of duplicity described by George Orwell in Nineteen Eighty-Four:

Day by day and almost minute by minute the past was brought up to date. In this way every prediction made by the Party could be shown by documentary evidence to have been correct; nor was any item of news, or any expression of opinion, which conflicted with the needs of the moment, even allowed to remain on record. All history was a palimpsest, scraped clean and reinscribed exactly as often as was necessary.

With an active search for disconfirming evidence, a judge describes her current beliefs using either an a priori analysis or bootstrapping. She then decides what information would be sufficient to change these beliefs, and conducts an active search for such data. Experimentation is especially important in this search for disconfirming evidence. The key question should be, “What information could possibly change my mind?” If you cannot answer this question, you cannot design an experiment that will contribute to your learning. If you have an exper-
A checklist was used for a cost-benefit analysis of the various types of forecasting methods. This provided the ratings in Exhibit 14-1. The major reasons behind these ratings were described.

Here are some things to keep in mind about costs:

- **DEVELOPMENTAL COSTS**
  - Armstrong's law: "There is no such thing as an easy job."

- **MAINTENANCE COSTS**
  - Murphy's law: "If something can go wrong it will."

- **OPERATING COSTS**
  - Parkinson's law: "Work expands to fill the time allotted."

For the assessment of uncertainty, judgmental methods offer a wide variety of inexpensive techniques, extrapolation methods are simple to use, and econometric and segmentation methods help in understanding the sources of the uncertainty.

The assessment of alternative futures is important because of the value of this information in planning. Econometric methods, scenarios, and role playing were highly recommended for assessing alternative futures.

When changes are large, as in long-range forecasting, it is important that the forecasting method contribute to learning. Extrapolation methods are of little value in this respect, but econometric methods are very helpful.

Unfortunately, we seldom use econometric methods for learning. Instead, we have great confidence in our ability to use judgmental forecasting methods and to learn from them. This confidence is unfounded! The unaided judge is an inefficient learner. (That includes me ... and possibly you also?) The primary reason is that people do not use disconfirming evidence. This problem is especially serious for people who view themselves as experts. They continue to use the same model even if conditions change. Experts should actively seek disconfirming evidence.
Even though people insist on using judgmental methods to forecast, much can be done to improve learning. The Brunswick lens model provides a convenient way to illustrate the types of feedback. A brief review of advice on improving judgmental learning is provided in Exhibit 14-3.

<table>
<thead>
<tr>
<th>Factors Relating to</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>Decomposition</td>
</tr>
<tr>
<td>Feedback</td>
<td>Fast feedback</td>
</tr>
<tr>
<td></td>
<td>Grouped data</td>
</tr>
<tr>
<td></td>
<td>Empirical studies</td>
</tr>
<tr>
<td>Feelings of expertise</td>
<td>Empirical studies</td>
</tr>
<tr>
<td></td>
<td>Avoidance of training programs</td>
</tr>
<tr>
<td>Active search</td>
<td>Experimentation</td>
</tr>
</tbody>
</table>