LONG-RANGE FORECASTING
From Crystal Ball to Computer
How should you evaluate forecasting models and forecasting processes? Part III discusses procedures for answering this question. Chapter 11 presents frameworks for the evaluation. Chapters 12 and 13 provide details for this framework; the former describes how to analyze the inputs to models, and the latter tells how to analyze the outputs from models.

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>Evaluation Schemes</td>
</tr>
<tr>
<td>12</td>
<td>Testing Inputs</td>
</tr>
<tr>
<td>13</td>
<td>Testing Outputs</td>
</tr>
</tbody>
</table>
LONG-RANGE FORECASTING
From Crystal Ball to Computer
Eleven
EVALUATION SCHEMES

Contents

Multiple Models .................. 298
When and What to Evaluate ...... 300
A Priori Framework ............... 301
A Posteriori Framework .......... 301
Inputs and Outputs ............... 302
Evaluating Acceptability .......... 303
Evaluating Quality ................ 306
  From the Real World to the Assumptions .......... 307
  From the Assumptions to the Model ................. 310
  From the Model to the Outputs .......... 311
  From the Outputs to the Real World ................. 312
Scoresheet for Evaluating Models .... 315
Evaluating the Forecasting System .......... 317
Summary .................................. 319
I don’t mean to deny that the evidence is in some ways very strong in favor of your theory, I only wish to point out that there are other theories possible.

— Sherlock Holmes

*Adventure of the Norwood Builder*

Arthur Conan Doyle

Sherlock Holmes is right this time. Even so, forecasters often ignore him! Much of the literature on forecasting methods reflects a “my method is best” position by authors. Few papers entertain the possibility that other methods (or theories or models) are superior. This chapter presents explicit frameworks for the evaluation of forecasting.

First I discuss the importance of using multiple models. Then I stress the need for an *a priori* framework for evaluating methods. This framework is recommended to provide a better evaluation and to improve the chances for implementation.

With respect to implementation, a distinction is made between the criteria of “acceptability” and “quality.” “Acceptability” refers to approval by those in the organization who would actually use the model; “quality” refers to the model’s ability to provide good outputs. A high-quality model that is not accepted is of no value, nor is a low-quality model that is accepted.

The chapter concludes with suggestions on evaluating the forecasting system. A checklist is provided to audit this system.

**MULTIPLE MODELS**

Statements such as “A mean absolute error of 20% represents a poor forecasting record” or “Uncertainties about the long-range future render any forecasts beyond 10 years worthless” are frequently encountered in papers on forecasting. Such statements, by themselves, say little; they remind me of that age-old question, “What’s the difference between an orange?” What is missing is a basis for comparison.

*The best way to evaluate the forecasting ability of a model is to compare it with other models.* Thus a mean absolute error of 20% may be very good indeed if the next best model has an error of 40%. And forecasts beyond 10 years may be useful if the proposed model does a better job than the implicit model that is currently being used. Nevertheless, if you insist on knowing the typical errors for a given situation, see MENTZER and COX [1984] and ASCHER [1978].
All aspects of model evaluation depend upon the comparison of alternative models. Which model has the most realistic assumptions? Which model provides the best predictions? Which model costs the least to develop? In other words, the beauty of a model is all relative; an illustration is provided in Exhibit 11-1.

The selection of a set of reasonable models is of critical importance. This step requires judgment by the analyst and by the clients. In most cases, one would expect the current model to be used as a basis for

Exhibit 11-1  THE BEAUTY OF A MODEL: The First Miss America (1921).

Source. © Miss America Pageant, permission for use granted.
Another guideline is that the models should stress variety; they should differ from one another to ensure that all reasonable approaches are examined. Still another guideline is to start with simple and low-cost models.

It is nice to appear to be reasonable and to present both sides of the issue. Nevertheless, I cannot think of anything worthwhile to say about research based on a single model. On the other hand, much published research on forecasting is done using a single model. Unfortunate.

WHEN AND WHAT TO EVALUATE

When should the plan for evaluation be developed, and what should it cover? The possibilities for "when" are *a priori* (before developing the forecasting models) and *a posteriori* (after developing the forecasting models).

The evaluation can focus on the process or on the content. Process relates to how the forecasting model will be evaluated and what will be the general framework for analysis. Content is more specific; it relates to an examination of the inputs and outputs for a given model. Although these categories overlap somewhat, they offer different perspectives.

Exhibit 11-2 presents four strategies for evaluation by considering the questions of what and when simultaneously. Box A, the *a priori* framework, is of particular concern in this chapter. Box B, the *a posteriori* framework, is a possibility. Although it is not recommended in this book, it is commonly used. Boxes C and D, the evaluation of inputs and outputs to a model, are considered only briefly in this chapter because they form the basis of Chapters 12 and 13.

<table>
<thead>
<tr>
<th>Process</th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>A priori framework</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Evaluate inputs</td>
<td>C</td>
<td>D</td>
</tr>
</tbody>
</table>

Exhibit 11-2 STRATEGIES FOR EVALUATION: WHAT AND WHEN
When and What to Evaluate

A Priori Framework

Chapter 3, on implementation, suggested that it is important to gain prior commitment from the key stakeholders. If this is not done, little meaningful change is likely to result. The prior commitment should include:

1. The models to be considered
2. The framework to be used
3. Specific criteria for items within the framework

Possible models were proposed in Part II of LRF; a framework for evaluation of models is provided in this chapter; and suggestions to help in the selection of criteria are presented here and in Chapters 12 and 13.

It requires much time and psychic energy to gain prior commitment—both for the forecaster and for the key stakeholders. It is uncomfortable; you feel that you are wasting time while the real work has not even begun. But this is a crucial step.

How do you know when you have gained prior commitment? Well, if you are not sure, you probably have not been successful here.

A Posteriori Framework

Forecasters who already know the right answer have no need for an a priori framework. They can do what many do—create a framework for evaluation after the forecasts have been obtained. This helps to ensure that the evaluation will not interfere with the desired solution. People who have worked on the evaluation of models know what I am saying here; others may think that I am being cynical. No, I am being realistic.

If the client has not committed himself to a framework for evaluation, he is free to contribute as he sees fit. Often the client sees fit to defend the existing methods. Or he may view himself as a problem finder—and find problems with your model or with your framework. This seems like a rational way to respond in such circumstances. (If you are not part of the solution, you are part of the problem.)

In response, the person proposing the forecasting model concludes that this is really a communication problem, so he spends a lot of time trying to explain the merits of his model.

At this point, each party acts as an advocate of his model. The cleverness of each researcher shows up as he tailors his framework to
demonstrate the point to be made. Typically, each framework considers only a part of the problem. An example of research in which the framework followed the development of the model is provided in the Ford-Chevrolet study described below. Of course, this is typical of academic research because seldom can one identify the clients in advance. That is why we have those strange "Comments on Doe," where Professor Oates points out the errors in Doe's paper; and the "Reply to Oates' Comments on Doe," where Doe points out that his paper was on apples whereas Oates was discussing oranges.

Evans (1959) used a linear discriminant analysis of psychological data to distinguish between buyers of Fords and Chevrolets. At least four papers criticizing Evans' study were published. Meanwhile, Evans wrote two replies to the critics. Admittedly some good blows were landed in all of this fighting. Still, one gets the impression that it is like watching the blind men feel the elephant; each researcher picked a framework to suit the point he wanted to make. It might have saved time and effort had Evans used a comprehensive framework for evaluation in his initial paper.

Inputs and Outputs

Most of us find it easier to deal with the specific rather than the general. Thus, instead of discussing a general framework for analysis, we prefer to jump into detailed questions about the inputs and outputs for a given model. This book suggests that inputs and outputs should be analyzed only after agreement has been reached on an a priori framework.

The analysis of inputs should come before the models have been developed. This helps to maintain people's interest in the inputs. It also keeps the researcher honest. When examining studies using industrial dynamics and exploratory regression analysis, I was told that the researchers often changed the inputs if they did not like the outputs. Forecasts become the center of attention once they are revealed. There is then less interest in the framework or in the inputs (except to change them to suit your biases). But if commitment was reached on the process, it is more likely that the forecasts will be accepted. The likelihood of rational decision making is increased if the decision makers have already considered what actions to take, given various forecasts.
If you do not reach commitment on the process, and if the clients are not prepared for alternative forecasts, you may have a problem. If you are lucky, you have a favorable forecast to report (i.e., one that requires no difficult changes for the organization). But what if the forecast brings bad news? If you are one of those “smart consultants,” like the Royal Meteorologist (Exhibit 11-3), you will make sure that the forecast sounds good. If you cannot make it sound good, you can use the fortune-teller's strategy and make it sound vague. So goes the dirty tricks department. If Royal Meteorologists and fortune-tellers know these tricks, then so do forecasting consultants.

EVALUATING ACCEPTABILITY

Acceptability can be evaluated by judging the model through the eyes of the user. In particular, the following should be examined:

1. Organizational value, that is, the user's perception of the value of the model to the organization.
2. Personal value, that is, the user's perception of the benefits (or costs) of the model to himself.

What is important here is not what actually exists—it is what clients believe to exist. Sometimes these perceptions match reality, but often they do not. Do not assume they are the same. Also, do not assume that the clients will perceive things in the same way that you do.

A written evaluation of acceptability should be made. The items in Exhibit 11-4 are typical of those that should be examined; also sug-
Exhibit 11-4  EVALUATING ACCEPTABILITY OF A MODEL

<table>
<thead>
<tr>
<th>User perceptions of organizational value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures (examples)</td>
</tr>
<tr>
<td>Profit</td>
</tr>
<tr>
<td>Employment</td>
</tr>
<tr>
<td>Customer service</td>
</tr>
<tr>
<td>Assessment techniques (examples)</td>
</tr>
<tr>
<td>Self-administered questionnaires</td>
</tr>
<tr>
<td>Nondirective interviews</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User perceptions of personal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measures (examples)</td>
</tr>
<tr>
<td>Security</td>
</tr>
<tr>
<td>Achievement</td>
</tr>
<tr>
<td>Power</td>
</tr>
<tr>
<td>Income</td>
</tr>
<tr>
<td>Assessment techniques (examples)</td>
</tr>
<tr>
<td>Direct questions</td>
</tr>
<tr>
<td>Projective questions</td>
</tr>
<tr>
<td>Group depth interviews</td>
</tr>
<tr>
<td>Role playing</td>
</tr>
<tr>
<td>Unobtrusive measures</td>
</tr>
</tbody>
</table>

gestions are provided on assessment. You should develop a checklist for your problem and apply it separately to each of the stakeholders. For an example of the assessment of alternative forecasting models, see GOMEZ-MEJIA, PAGE, and TORNOW [1982].

The assessment of organizational value is generally easier than the assessment of personal value, because the former is an acceptable thing to talk about. A major problem, however, is that people in organizations view a proposal in terms of its impact upon their particular units within the organization. Furthermore, it is generally taboo to question the goals of a unit. As a result, perceptions of the value of a proposed method often vary substantially by group (e.g., labor vs. management, production vs. marketing, stockholder vs. customer, line vs. staff).

Nondirective interviewing can be useful in getting people to express
opinions without being constrained by these organizational taboos. So can self-administered questionnaires. Certainly the replies by each respondent must be kept confidential; thus it helps to have the surveys administered by people outside the organization.

The assessment of personal value is affected by a client's desire to answer so that he appear in a favorable light. For this reason, the assessment techniques in Exhibit 11-4 emphasize indirect approaches. (These techniques were discussed in Chapter 6.)

Beyond the use of checklists and a written evaluation of the acceptability of proposed forecasting models, little more can be said. Just use Gerstenfeld's law of trying.

In the following, I describe the approach used to evaluate the acceptability of a forecasting model that was being used by a large U.S. corporation. The company, Drinkit Corporation, had asked us to evaluate the model. The names have been disguised. The disguise fooled some people, who incorrectly thought their company was being described! Apparently, this description seems typical for organizations.

Armstrong and Shapiro (1974) examined the FAITH models, which are used by some of the largest corporations in the United States. Their advocates claim that the models are useful for predicting changes in market share when changes are made in prices, product line, or advertising. To evaluate the acceptability of these models, we interviewed key user groups, including marketing research, sales, product management, marketing management, and general management in the company. Individual and group interviews were used, and they were primarily nondirective. Two interviewers sat in on each session and took notes separately so that the reliability of the information could be improved. Highlights of these interviews are summarized here.

**Perceived Organizational Value.** Perceptions of the value of FAITH were mixed. The top two levels of management in Drinkit thought that FAITH provided better predictions than those currently made by the product managers. The middle and lower levels felt just the opposite, partially because they had confidence in their own judgment. Comments by the product managers included: "FAITH can't be wrong; if their estimates are off, the advocates claim it is because the input data were not right."
Perceived Personal Value

1. Opinion was split by organizational level: the higher levels thought that FAITH was valuable because it gave them more control over decision making; the middle and lower levels felt that FAITH was another attempt by higher managements to reduce subordinates' influence.

2. The low level of acceptability of FAITH also showed up in unobtrusive measures. For example, when the middle and lower levels of Drinkit management met for an all-day planning session, not one reference was made to FAITH—yet FAITH was then regarded by top management as the foremost quantitative planning model in the company.

3. Because the marketing managers were told by top management to use the FAITH models, middle- and lower-level managers would change the inputs to the FAITH model until they found a result that agreed with their own decision. This version was then presented to top management. In other words, managers were misusing the models.

The analysis of acceptability is often crude; however, the problems are sometimes of such magnitude that a precise examination is not required. In the case of Drinkit, it quickly became apparent that the major user group, the product managers, had little faith in FAITH.

EVALUATING QUALITY

Quality can be evaluated by examining the inputs (assumptions) to the model, or by examining the outputs (forecasts) from the model. The testing of inputs has been more popular. Some, such as Machlup (1955), have gone so far as to imply that the testing of inputs is the only worthwhile way to test models.

An equally unreasonable position was taken by Friedman (1953), who claimed that the testing of outputs is the only useful approach for testing models. (Nagel, 1963, criticized Friedman's position.)

It is better to test both the inputs and the outputs. The primary reason for testing inputs it to learn how to improve a given model. The major reason for testing outputs is to select the best model(s). Naturally, there is some overlap; tests of inputs may demonstrate that one
model is clearly inferior to another, and tests of outputs may provide clues to improve the model.

These ideas on testing inputs and outputs can be translated into somewhat more operational terms by examining how the forecasting model relates to the real world. These relationships are illustrated in Exhibit 11-5. The right-hand side of the exhibit represents the inputs to the model: “Are the assumptions reasonable?” and “Does the model follow from the assumptions?” The left-hand side represents the outputs from the model: “Can the outputs be replicated?” and “Do the benefits from the model exceed the costs?” In the following, these four stages are discussed.

From the Real World to the Assumptions

Are the assumptions behind the model reasonable? To assess this, you should search for evidence relating directly to these assumptions. Below, ideas are presented on how to search for this evidence.

Exhibit 11-5 STAGES IN ANALYZING THE QUALITY OF A MODEL

One key decision is whether to use objective (empirical) or subjective data to test assumptions. Clearly, objective data, such as those obtained from experiments, are preferred. If objective data are not available, then you must use subjective data. For the assessment of acceptability you might use both objective and subjective data.

Another key decision is whether to utilize experts from inside the organization or from outside it. Outside experts are less subject to bias; on the other hand, their evidence may be less relevant to the specific problem faced by the organization. Bias is more serious for subjective data, although it can also occur for objective data. The issue of relevance is particularly important for objective data because it is important to ensure that these data were generated in a situation representative of the problem at hand.

These ideas about searching for objective vs. subjective evidence, and using inside vs. outside experts, are summarized in Exhibit 11-6. Within that exhibit, some specific techniques are proposed. The preferred approach (source $a$) is to use objective evidence generated from experiments for the problem at hand. Next best (source $b$) is to use objective empirical evidence developed by others. If this fails, you could use subjective impressions of people within the organization (source $c$). Alternatively you might turn to subjective evidence from unbiased outside experts (source $d$); this could come from papers published by experts or from hired consultants.

Unfortunately, it is frequently necessary to resort to subjective evidence from inside experts. The power of this approach can be improved by use of the following techniques:

<table>
<thead>
<tr>
<th>Exhibit 11-6 SOURCES OF EVIDENCE ON ASSUMPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Objective</strong></td>
</tr>
<tr>
<td>$a$</td>
</tr>
<tr>
<td>Experiments</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$b$</td>
</tr>
<tr>
<td>Empirical studies</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Surveys of the people in an organization provide a fast, inexpensive way to determine whether they regard the assumptions for a model as reasonable. It is preferable that the surveys be self-administered and anonymous. The questions should provide a choice among assumptions rather than absolute answers. For example, an assumption may be stated and its exact opposite used as an alternative. The respondent could indicate the degree of agreement with each of these extremes.

Unobtrusive measures may be helpful for emotional issues. Sometimes the organization trains people to talk in certain ways, yet they act differently. Is it possible to observe how people act and thus infer their beliefs? For example, people in a personnel department may claim that their predictions of the probable success of job applicants do not depend upon age, race, sex, or religion. Do they act that way?

Indirect bootstrapping is a logical extension of unobtrusive measures to infer the assumptions that were used to make predictions. For the personnel example, variables on age, race, sex, and religion could be examined.

An example from our study of the FAITH models illustrates how some of these ideas can be used:

In Armstrong and Shapiro's (1974) study of the FAITH models, the advocates of FAITH based their claims on the reasonableness of the assumptions on face validity, which in this case meant that the assumptions looked reasonable to them. Strangely, however, they asked the clients to suspend their beliefs when evaluating the assumptions. (This is true. You were probably wondering where we got the name "FAITH models.") The following three assumptions were among those examined:

1. The models assumed that the switching between any two brands of beverages was equal in both directions; that is, the number of customers switching from Mother Fletcher's brand to Bestbuy is the same as the number switching from Bestbuy to Mother Fletcher's for a given period of time. No experimental evidence on this was available from within the company despite the fact that it had been evaluating FAITH over a six-year period (source a in Exhibit 11-6).

2. It was assumed that the forecasts were not sensitive to mea-
measurement error in the data. We tested this using FAITH predictions for one of Drinkit’s beverages, which was referred to as Kola. Predictions for each of 15 time periods were compared with actual data when different starting points were selected. When period 1 was used to obtain estimates for the FAITH model, the MAPE for periods 2 through 16 was 4.9%. When period 3 was used for estimation, the MAPE for the remaining 15 periods (i.e., periods 1, 2, and 4 through 16) was 2.7%. In all, six different starting points were examined, and the MAPE varied from 2.7% in the best case to 12.7% in the worst case. The implication is that forecast accuracy was highly sensitive to measurement error. (Source a in Exhibit 11-6.)

3. Published empirical evidence from Morrison (1966) suggested that the assumption about equal switching in each direction was unreasonable. (Source b in Exhibit 11-6.)

4. It was assumed that the brand of beverage purchased by a consumer was unrelated to the brand previously purchased by that consumer. Interviews with Drinkit’s product managers indicated that this assumption was unreasonable. (Source c in Exhibit 11-6.)

5. The assumptions looked unreasonable to us as consultants. (Source d in Exhibit 11-6.)

In general, the users’ assumptions conflicted with those employed in the FAITH models. Furthermore, we were unable to locate a single client who claimed to have an adequate understanding of the models. Typical comments were: “No one can explain FAITH to me”; “I don’t know how FAITH works.”

From the Assumptions to the Model

This stage of analysis involves an examination of the logical structure of the forecasting model. Here it is important that the structure of the model be fully disclosed. Unfortunately, claims of “competitive secrecy” are often used to rule out full disclosure. Such claims are sometimes used by charlatans.

A more sophisticated alternative used by some charlatans is to provide full disclosure, but to use complex procedures and difficult languages so that the clients cannot understand what is being said. Bafflegab puts the clients in an awkward position. The clients are apparently
being given all of the information about the model’s structure, and the advocate is spending much time with them trying to explain the assumptions; are they going to admit that they are so stupid that they cannot understand the model? The easy way out is to nod in agreement and to hope the advocate knows what he is talking about.

Bafflegab can be used to divert the client’s attention from other stages of analysis. They exhaust themselves trying to understand the model, finally concluding that it is too complex for them and that they must trust the model builder.

Complexity is no virtue in forecasting, as was discussed in Part II. Thus there is no excuse for bafflegab, even if practiced by well-intentioned and pleasant people. One should insist that the change agent’s duty is to explain the model so that the client can understand it. Good tests of understanding are for the client to:

1. be able to explain the structure of the model to someone else, and to
2. be able to explain the structure of the model in written form so that it makes sense to you.

The change agent should help to ensure that an understanding is reached. This brings us back to acceptability.

The advocates of FAITH used bafflegab. Although the model was not shown to be inconsistent, an immense amount of effort was required to work through the logical structure:

In Armstrong and Shapiro (1974), a detailed examination was made of the logical structure of the model, which the advocates referred to as FAITH-DYNAMICS. This examination was time consuming because of the model’s complexity and because of poor documentation (we had to meet with the advocates for much of the explanation). We did not find any logical inconsistencies in the model. We then contacted other researchers who had made independent evaluations of this model; they also had detected no inconsistencies but, like us, had found the model to be obscure.

From the Model to the Outputs

Stage 3 is a routine auditing step. Given the model and the data, is it possible to replicate the output? You take a sample of the data that
were used and enter it into the model that the advocates used. The procedure is analogous to a financial audit.

It is preferable to do the calculations by hand, using the model that you described in your own words.

This stage of analysis is often overlooked; we trust the model builders to be competent. But this stage has the following advantages:

1. It helps the clients to ensure that they understand the model.
2. It provides an additional check against errors.
3. Cheating is unusual, and this procedure helps to keep it unusual.

The point on honesty should not be overlooked. Cyril Burt, who is dead now, had been described as "one of the world's great psychologists." His main claim to fame had been his study of the IQs of identical twins. Strangely, as Burt published studies in 1955, 1958, and 1966, his sample sizes of identical twins increased from 21 to "over 30" and then to 53 pairs, yet the correlation between the IQ scores for identical twins was .771 in all three cases. One hypothesis is that he was cheating. (For a description of this case see Wade (1976).) Nevertheless, do not accuse anyone of cheating. Failures to replicate may arise in many ways, such as by mistakes. Speaking of replication, we were unable to replicate the FAITH results for Drinkit:

Armstrong and Shapiro (1974) used the same data that were used by FAITH advocates to predict for 15 periods of Kola. For 12 of the 15 periods, the predictions by the FAITH advocates were more accurate than those obtained by us, using their data and their model. Their average error was half the error that we found. The advocates were unable to explain this discrepancy.

From the Outputs to the Real World

The analysis of outputs is the most important stage in assessing the quality of a model. Rather than attack the problem directly, you should break it into parts by means of a cost-benefit framework. The cost-benefit framework helps to make the examination more systematic and explicit.

Given a set of replicable models with realistic assumptions, the general procedure for a cost-benefit analysis is as follows:
1. List the potential benefits for each model.
2. List the potential costs for each model.
3. Weigh benefits against costs for each model.
4. Select the model with the most favorable cost-benefit score.

Although this list is simple from a conceptual viewpoint, it is hard to implement. Gerstenfeld’s law of trying is recommended. Also recommended is Exhibit 11-7, which can be used as a checklist to ensure that all costs and benefits have been examined. It describes three cost items: initial development costs, maintenance costs (to keep the model up to date), and operating costs (time and dollars spent by users to obtain the forecasts). There are four benefits: improved accuracy of the forecasts, assessment of uncertainty (how much confidence should we place on the forecast?), assessment of alternative futures (including changes in the environment or changes in the organization’s policies), and the contribution of the model to learning (will the forecasting model improve over time?).

The selection of alternative models is of particular concern in the cost-benefit analysis. Although it is best to have explicit alternatives, two procedures can be used with implicit alternatives:

- The test of large changes
- The Turing test

The first procedure involves the use of extreme inputs to test the generality of the model. This is especially useful if there are extreme inputs whose outcomes would be obvious. The key question is whether the model provides reasonable forecasts for large changes.

The Turing test (Turing, 1950) involves comparisons of the outputs from various models. This can be done even with forecasts from implicit models. The question is whether a panel of experts can distinguish differences in the reasonableness of outputs from different models. This procedure is often used in the research on expert systems.

<table>
<thead>
<tr>
<th>Costs</th>
<th>Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development</td>
<td>Forecast accuracy</td>
</tr>
<tr>
<td>Maintenance</td>
<td>Assessment of uncertainty</td>
</tr>
<tr>
<td>Operating</td>
<td>Assessment of alternative futures</td>
</tr>
<tr>
<td></td>
<td>Learning</td>
</tr>
</tbody>
</table>
When using explicit alternatives, you should include the model currently used. This model will serve as a benchmark for comparisons.

Another explicit alternative should be based on simplicity. In Chapter 10, bootstrapping was proposed as a way to translate complex judgmental forecasting into a simple quantitative model. Bootstrapping can also be used to translate a complex quantitative model into a simple quantitative model. This procedure is illustrated in our study of the FAITH models, along with some of the steps used in the cost-benefit analysis.

In Armstrong and Shapiro (1974), the assessment of costs was relatively easy. Drinkit Corporation paid about $60,000 per year to the FAITH Corporation, and an additional $40,000 was estimated for the time spent by Drinkit personnel (in 1973 dollars). These costs were constant from year to year because there was a continuing need for expert assistance. On the benefit side, FAITH provided no assessment of uncertainty. FAITH advocates claimed, however, that their models were useful for improving accuracy, examining alternative futures, and learning. No empirical test had been made of these claims in six years of use by Drinkit.

Because Drinkit was especially interested in forecast accuracy, we examined forecasts using large changes; an examination was also made of the accuracy of alternative models.

**Large Changes.** The FAITH models multiplied the following variables to obtain the best advertising level: industry volume \( \times \) unit margin for brand \( \times \) brand market share \( \times \) competition's market share \( \times \) brand switching. Under this model, if industry volume doubled, Drinkit should spend twice as much on advertising. This did not seem reasonable compared to the implicit forecasting model used by the product managers at Drinkit.

**Alternative Models.** Drinkit was unable to provide any cases where FAITH predictions were compared with the managers' judgmental predictions. We compared the FAITH forecasts with those from a simple quantitative model obtained by bootstrapping the FAITH model. That is, regression analysis was used, with the dependent variable based on FAITH forecasts, and the causal variables were those used by the FAITH model. This yielded a SON-OF-FAITH model. The advertising version of this model was
\[ Y = 20.7 + 0.6X \]

where \( Y \) = market share (predicted by FAITH model)
\( X \) = advertising dollars (divided by 1 million)

The SON-OF-FAITH model explained 98% of the variance in FAITH predictions.

Because SON-OF-FAITH required FAITH predictions, the model was recalibrated using actual data; four periods of Kola data were used. Forecasts were then made for periods 5 through 16, and the forecasts were compared with actual data. The MAPE for these predictions was 2.7% compared with a MAPE of 3% for predictions by the FAITH model. The superiority of this simple econometric model was also found when different starting values were used, although the differences were not statistically significant.

**SCORESHEET FOR EVALUATING MODELS**

The final step in the framework for analysis is to summarize the models according to the criteria and weights initially agreed upon. If you did a good job in gaining commitment, this is a routine step; you plug in the results, and the decision falls right out. If you did not reach commitment, there are at least three possibilities:

1. Everyone thinks the problem is unimportant.
2. You got the “right” answer (i.e., the answer the decision makers already wanted).
3. You are willing to create a confrontation to gain acceptance of your findings.

Armstrong and Shapiro describe the use of the scoresheet in the FAITH study.

One of the bad things about the Armstrong and Shapiro (1974) study was that we were not able to gain commitment. Nevertheless, we completed the analysis and obtained the results shown in Exhibit 11-8. The FAITH models were compared to the subjective process currently used by managers and also to the simple
econometric model. The entries in Exhibit 11-8 are based upon our judgments after conducting various tests, some of which were described previously. Each exhibit entry is a rating of FAITH. Thus, for user perception of quality, FAITH rated as poor relative to management’s judgment. The simple econometric model clearly dominated the FAITH model; it would have been superior to FAITH under any weighting scheme. Management’s judgment was also superior in light of the problems expected with the acceptability of FAITH.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Acceptability</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>User perception of quality</td>
<td>Stage 1: Reasonable assumptions</td>
</tr>
<tr>
<td></td>
<td>User perception of personal value</td>
<td>Stage 2: Logical structure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 3: Audit of outputs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Stage 4: Cost-benefit analysis</td>
</tr>
<tr>
<td>Ratings of FAITH vs.</td>
<td>Poor</td>
<td>Poor</td>
</tr>
<tr>
<td>Management’s Judgment</td>
<td>Poor</td>
<td>About the same</td>
</tr>
<tr>
<td>Simple Econometric Model</td>
<td>Unknown</td>
<td>About the same</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poor</td>
</tr>
</tbody>
</table>


Our framework led to questions that had not been examined during the six years of evaluation by the Drinkit Company. These questions, in turn, led to conclusions that differed sharply from
those that had been reached previously by Drinkit. Ah, but the lack of commitment! We were able to predict Drinkit's decision: They kept the FAITH.

EVALUATING THE FORECASTING SYSTEM

The preceding part of this chapter was oriented to an evaluation of forecasting models. For an overall examination of the forecasting system in an organization, I developed the “Forecasting Audit Checklist” (Exhibit 11-9). This checklist, based heavily upon the research in LRF, was originally presented in ARMSTRONG [1982b]. It addresses issues related to forecasting methods, assumptions and data, uncertainty, and costs.

The items on the checklist are worded in such a way that a “YES” is the desired response. My students and I have used the checklist to audit numerous organizations. In the following, I present some impressions from these applications. The numbers correspond with the items on the checklist:

1. Top management should be involved primarily in providing inputs to the forecasting system. But their involvement in the forecasting methods is expensive and is likely to lead to bias. In practice, most of the organizations we examined did not keep the forecasting methods independent of top management.

2. Objective methods can often reduce costs (e.g., in inventory control) or improve accuracy. For example, bootstrapping provides gains relative to unaided judgment, and econometric models provide further gains in accuracy.

3. Evidence that structured techniques can improve the accuracy of judgmental forecasts is impressive. Nevertheless, this item was often ignored by the organizations we examined.

4. Only a moderate amount of expertise is required to forecast change, though expertise is important for estimating current status. Some firms could save money by remembering this point.

5. The majority of the organizations based their forecast on the single “best” method. Few combined forecasts.

6. As firms move toward the use of objective methods, the problems involved with understanding seem to increase.

7. One of the most frequently violated guidelines was that com-
### Exhibit 11-9 FORECASTING AUDIT CHECKLIST

<table>
<thead>
<tr>
<th>Topic Areas</th>
<th>No</th>
<th>?</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FORECASTING METHODS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Forecast independent of top management?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Forecast used objective methods?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Structured techniques used to obtain judgments?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Least expensive experts used?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. More than one method used to obtain forecasts?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Users understand the forecasting methods?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Forecasts free of judgmental revisions?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Separate documents prepared for plans and forecasts?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>ASSUMPTIONS AND DATA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Ample budget for analysis and presentation of data?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Central data bank exists?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Least expensive macroeconomic forecasts used?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>UNCERTAINTY</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Upper and lower bounds provided?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Quantitative analysis of previous accuracy?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Forecasts prepared for alternative futures?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15. Arguments listed against each forecast?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>COSTS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16. Amount spent on forecasting reasonable?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
panies frequently made judgmental revisions of the objective forecasts of change.

8. Separate documents for plans and forecasts can help to solve difficulties associated with keeping the forecast independent of top management (item 1) and with avoiding judgmental revisions (item 7). It can also aid in assessing the accuracy of the forecast (item 13).

9. The budget for analysis and presentation of data often suffers relative to the costs for data collection. This item argues against that automatic response “we need more data”; instead, more emphasis should be given to the analysis and presentation of available data.

10. Many organizations report that the information needed to make forecasts is not available in a central location (e.g., a computer file).

11. Given the lack of evidence that one econometric model is more accurate than another, the decision on which econometric service to use should be based on cost or other criteria, such as ease of understanding.

12. Most organizations reported problems with the assessment of uncertainty. Forecasts are commonly presented (often with many significant digits!) with no mention of upper and lower confidence limits.

13. Many firms have no systematic procedures to track the accuracy of their forecasting methods.

14. Few firms reported that they prepared forecasts for alternative futures.

15. It is not common for an organization to explicitly list the arguments against each forecast.

16. Few firms kept track of the amount spent on forecasting. Nevertheless, the general amount thought spent on forecasting was reasonable.

SUMMARY

To evaluate forecasting models, compare them with alternative models. Of particular interest are comparisons with models that represent current practice, and models that are inexpensive and simple.

The scheme for evaluation should be developed before the data are analyzed. It is especially important that commitment be reached on the process used to evaluate the models. Little is gained by creating an evaluation framework after the forecasting models have been developed. Some attention should also be given to the content of the forecasting model, that is, to the inputs and outputs.

Acceptability and quality should each be explicitly evaluated for a
proposed forecasting model. The model should score well in each area before it is adopted.

Acceptability can be examined by measuring the clients' perceptions of the value of a model not only to the organization, but to themselves. A checklist of methods was presented in Exhibit 11-4 for assessing these perceptions.

Quality is assessed by the four-stage approach outlined in Exhibit 11-5. This examines the reasonableness of the inputs, the logical consistency of the model, the replicability of outputs from the model, and a cost-benefit analysis of the outputs.

An explicit scoresheet was suggested for bringing together the results of the evaluations of acceptability and quality of a model (Exhibit 11-8).

Finally, a forecasting audit checklist was proposed to aid in the evaluation of a forecasting system (Exhibit 11-9).