University of Pennsylvania
ScholarlyCommons

Marketing Papers

1-1-2001

Evaluating Forecasting Methods

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Suggested Citation:
Armstrong, J.S. Evaluating Forecasting Methods. In *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Ed. J. Scott Armstrong). Kluwer, 2001.

Publisher URL: http://www.springer.com/business+%26+management/business+for+professionals/book/978-0-7923-7930-0

This paper is posted at ScholarlyCommons. http://repository.upenn.edu/marketing_papers/146
For more information, please contact repository@pobox.upenn.edu.

Evaluating Forecasting Methods

Abstract
Ideally, forecasting methods should be evaluated in the situations for which they will be used. Underlying the evaluation procedure is the need to test methods against reasonable alternatives. Evaluation consists of four steps: testing assumptions, testing data and methods, replicating outputs, and assessing outputs. Most principles for testing forecasting methods are based on commonly accepted methodological procedures, such as to prespecify criteria or to obtain a large sample of forecast errors. However, forecasters often violate such principles, even in academic studies. Some principles might be surprising, such as do not use R-square, do not use Mean Square Error, and do not use the within-sample fit of the model to select the most accurate time-series model. A checklist of 32 principles is provided to help in systematically evaluating forecasting methods.

Disciplines
Business | Marketing

Comments
Suggested Citation:
Armstrong, J.S. Evaluating Forecasting Methods. In *Principles of Forecasting: A Handbook for Researchers and Practitioners* (Ed. J. Scott Armstrong). Kluwer, 2001.

Publisher URL: http://www.springer.com/business+%26+management/business+for+professionals/book/978-0-7923-7930-0

This book chapter is available at ScholarlyCommons: http://repository.upenn.edu/marketing_papers/146

Principles of Forecasting: A Handbook for Researchers and Practitioners,
J. Scott Armstrong (ed.): Norwell, MA: Kluwer Academic Publishers, 2001.

Evaluating Forecasting Methods

J. Scott Armstrong
The Wharton School
University of Pennsylvania

ABSTRACT
Ideally, forecasting methods should be evaluated in the situations for which they will be used. Underlying the evaluation procedure is the need to test methods against reasonable alternatives. Evaluation consists of four steps: testing assumptions, testing data and methods, replicating outputs, and assessing outputs. Most principles for testing forecasting methods are based on commonly accepted methodological procedures, such as to prespecify criteria or to obtain a large sample of forecast errors. However, forecasters often violate such principles, even in academic studies. Some principles might be surprising, such as do not use R-square, do not use Mean Square Error, and do not use the within-sample fit of the model to select the most accurate time-series model. A checklist of 32 principles is provided to help in systematically evaluating forecasting methods.

Keywords: Backcasting, benchmarks, competing hypotheses, concurrent validity, construct validity, disconfirming evidence, domain knowledge, error measures, face validity, goodness of fit, jackknife validation, outliers, predictive validity, replication, statistical significance, and successive updating.

Principles have been developed to guide forecasters in selecting a forecasting method (Armstrong 2001b). However, decision makers may be unwilling to generalize from prior research, believing that their situation is different. Or prior research may have revealed a number of relevant methods and one would like to narrow the field. This calls for systematic testing in the situation in which the forecasts will be used or in a closely related situation. This paper discusses procedures for such testing.

REASONABLE ALTERNATIVES

The value of a method should be judged by comparing it with other methods. Sherlock Holmes commented on this issue: “I don’t mean to deny that the evidence is in some ways very strongly in favor of your theory. I only wish to point out that there are other theories possible.” (A. C. Doyle, *Adventure of the Norwood Builder*, 1905)

- **Compare a proposed method against reasonable alternatives**

In 1890, Chamberlin (reprinted in Chamberlin 1965) claimed that sciences that use multiple competing hypotheses progress more rapidly than those that do not. Empirical research since then (summarized by Armstrong, Brodie and Parsons 2001) supports this claim.

Competing methods should be reasonable. How can one judge this? Selecting reasonable alternative methods requires knowledge of the forecasting literature. Textbooks provide descriptions of forecasting methods and experts can provide advice about them.

Sometimes alternative methods are used but they are not reasonable. Gurbaxani and Mendelson (1990) compared their forecasting model, which used 28 time-series observations, against an alternative model that used only two observations. In addition, they used recent data to calibrate their preferred model and older data to calibrate the competing model. For another example, consider Focus Forecasting, which is popular among practitioners. It

compares alternative extrapolation methods, but their methods do not represent the state of the art (Gardner and Anderson 1997).

Simple methods may be reasonable. According to evidence summarized in Armstrong (1984), simple extrapolation models (such as the naive model that “things will not change”) are often accurate. Schnaars (1984), for example, used extrapolation methods to produce annual forecasts for five years ahead for 98 annual series representing sales of consumer products; the naive forecast was as accurate as any of the other five extrapolation methods he used.

Selecting reasonable alternative methods for cross-sectional data is more difficult than it is for time series, but the 'base rate’ provides a good starting point. Base rates describe typical behavior. For example, a model’s forecast that a new product will be successful might be compared with a base rate, such as the percentage of new products in this category that have been successful over the past two decades.

The current method is usually a reasonable alternative. For example, the current process could serve as a benchmark when examining the value of a new method to predict the outcomes of negotiations. Or, if one proposes a new procedure for selecting executives, the current procedure (which is likely to be a series of unstructured group meetings) should be used as one of the alternative methods.

Researchers often fail to use reasonable alternative methods in their tests. Armstrong (1979), in an examination of empirical papers published in *Management Science* from 1955 through 1976, found that only 22% used multiple competing hypotheses. Armstrong, Brodie and Parsons (2001), in a study of five leading marketing journals from 1984 through 1999, found that less than 15% of the empirical studies examined reasonable alternative hypotheses.

Relative to other areas of management science, forecasting appears to have a decent record with respect to comparing alternative approaches. I coded a sample of 105 empirical papers published by the *Journal of Forecasting* and the *International Journal of Forecasting* and found that 58% of these examined reasonable competing hypotheses (Armstrong 1988).

FRAMEWORK FOR EVALUATING FORECASTING METHODS

One can evaluate a forecasting method by examining its inputs or its outputs. While this might seem obvious, it has long been the subject of debate. Friedman (1953) claimed that testing outputs is the only useful approach to evaluating methods. Nagel (1963) criticized Friedman’s position, and Machlup (1955) claimed that testing inputs is the only worthwhile way to test methods.

It seems reasonable to test both inputs and outputs. The primary reasons for testing inputs are to learn how to improve a given model and, in the case of causal models, to better assess the effects of policy changes. The major reasons for testing outputs are to select the best models and to assess uncertainty. However, tests of inputs may show that one model is inferior to another, and tests of outputs may provide ideas about how to improve the model.

With respect to inputs, there are two key questions: Are the assumptions reasonable? Are the proper methods and data used? With respect to outputs, the questions are: Can the outputs be replicated? How can one assess the outputs of the model?

TESTING ASSUMPTIONS

Bretschneider et al. (1989) obtained results consistent with the need to test assumptions. They found that states in the U.S. that used a formal process to question assumptions obtained more accurate forecasts of government revenues.

Use objective tests of assumptions.

Objective data, such as those obtained from experiments, are preferable for testing assumptions. If objective data are not available, then you can use subjective information. This can come from experts inside or outside the organization. Inside experts have more relevant information than outside experts, but they tend to be more biased.

Surveys of experts can help determine whether they regard the assumptions for a given forecasting model as reasonable. The respondents could be asked to choose among alternative assumptions. The surveys should be self-administered and anonymous.

An example illustrates some of these ideas. In Armstrong and Shapiro's (1974) study of the FAITH models (fictitious name), the following assumptions were used in predicting market share given various levels of advertising:

(a) Switching between any two brands is equal in both directions; that is, the number of customers switching from Brand A to Brand B is the same as the number switching from Brand B to Brand A for a given period. The consultants had no evidence to support their assumption. Published empirical evidence by outside experts showed that this assumption was probably false.

(b) The brand of beverage a consumer purchased is unrelated to the brand purchased previously by that consumer. In interviews, the company's product managers said that this assumption was unreasonable.

In general, then, assumptions of the FAITH models were unsupported and incorrect.

- **Test assumptions for construct validity.**

To test construct validity, one compares different approaches to estimating a given variable or relationship. You can test construct validity to identify unreasonable assumptions. It is especially useful for testing econometric models. For example, I developed an econometric model to forecast photographic sales in international markets. The estimates of parameters from various sources were all in rough agreement, thus providing evidence of construct validity. For income elasticity, prior research using time series yielded an estimate of 1.3, household survey data provided an estimate of 1.5, a cross-sectional analysis of countries yielded an estimate of 0.9, and a longitudinal analysis across countries produced an elasticity of 1.6. Consequently, I used the average income elasticity of 1.3 (Armstrong 1985, p. 330).

Assumptions are especially important when models are based on causality. Stephan (1978), who testified before the U.S. Supreme Court in Brown vs. Board of Education, concluded years later that he had made an incorrect forecast about the effects of bussing to achieve racial diversity in schools. He had assumed that the school's environment would not affect the forecast. As it turned out, integration does not improve social interactions and attitudes when people are put into competitive situations, such as public schools.

Describe conditions of the forecasting problem.

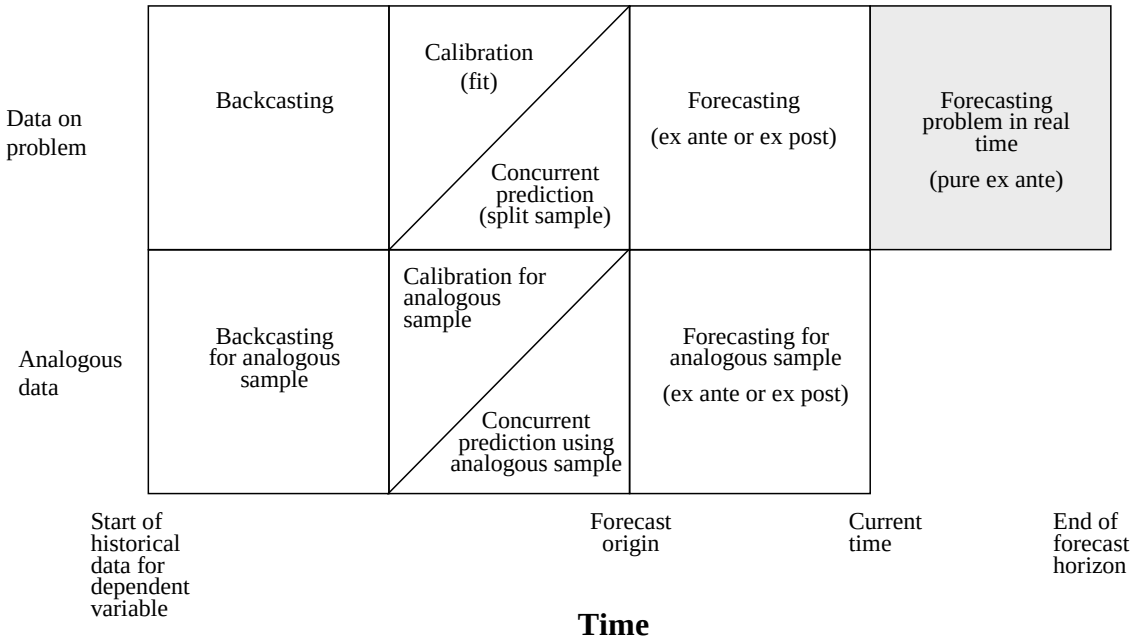
Forecasters need an accurate description of the conditions underlying a forecasting problem in order to develop generalizations. Knowledge about conditions will also help practitioners match prior research to their situation. However, one problem has been to have researchers agree on ways to describe conditions. Some conditions are obvious, such as the forecast horizon or the period of the observations (e.g., annual or monthly). Armstrong, Adya and Collopy (2001) list 28 conditions that can be used to describe time series.

While the M-competition contained information on some key conditions, such as the length of the forecast horizon and the period of the data, its descriptions were vague (Makridakis et al. 1982). This makes it difficult for practitioners to relate these studies to their own situations. The M-competition compounded the problems by summarizing across data where conditions obviously differed. For example, monthly, quarterly, and annual forecast errors were combined in some analyses.

- **Design forecasting tests to match the forecasting problem.**

Forecasting methods should be tested in situations that resemble the actual situation. The basic assumption here is that the closer the correspondence between the predictive test and the criteria, the better the predictive validity. The possibilities are illustrated in Exhibit I. Ideally, one would prefer to test pure ex ante forecast validity. For practical reasons, this may be impossible to test. For example, you might not be able to wait that long. As a result, you might turn to alternative procedures that are discussed below.

Exhibit 1
Validation Matrix: Data for Testing Forecast Validity



Methods should be tested using data that are relevant to the problem. Thus, in finding the best method to forecast automobile sales in a country, it would be desirable to use automobile data from that country. However, you could also use analogous data from other countries.

Researchers in organizational behavior often depend on analogous data. Consider the problem of forecasting which individuals are likely to be successful employees. Past performance in the same job would be a good indicator. However, if that is not available, one should try to find or create a similar situation. For example, researchers have developed realistic job-sample tests to determine how candidates are likely to perform on the job as shown by the following two studies.

Smith (1976) used a “trainability test” for dental students. He identified the crucial elements of a job, found skills and knowledge that can be imparted in a short time, and then created a test that was complex enough to allow the typical job applicant to make a number of errors. Candidates who learned the job more easily on this test were later found to be more successful on the job.

In another study that used realistic analogous data, Dalessio (1994) examined the predictive validity of the responses of new insurance agents. The agents saw video-tapes that presented realistic but unresolved selling situations. In the situations taped, prospects gave their objections to agents when asking for appointments, closing sales, and requesting referrals. Following each video-tape, the new agents were given multiple choice questions asking them what they would do if they were the agent in the situation. The test was administered in 14 insurance companies to 677 agents who had worked less than three months on the job. This occurred after the normal selection procedures had been used to hire the agents, so it was an additional hurdle in the selection process. Because the overall retention rate for the companies in this sample was already about 10% better than the industry average, this was a difficult situation in which to show improvement. Dalessio randomly divided the sample into three parts to

test the validity of the video test. He used each part successively as a holdout validation sample, basing the item-scoring weights on the rest of the sample. The video-based selection test proved to be predictive of turnover. The average retention for those who scored in the top quarter of the test was 78%, while for the lower-scoring quarters the retention rates were 66%, 62%, and 59%, respectively. In other words, the analogous data provided a useful test of forecast validity.

It is best to use holdout data for the future. Lacking that, one can use data for the current period (concurrent validity). Another, though seldom used possibility, is to forecast backwards in time (backcast validity). Backcasting is useful for testing econometric models when adequate data on causal variables exist only for recent years. You could backcast earlier values, say for 1990 to 1980, by using more recent data, say from 2001 to 1991. These backcasts could then be compared with the actual values for the dependent variable. This is most appropriate when the causal effects have been captured within the time interval, so annual data are likely to be relevant. For shorter periods, one might be concerned about the transition pattern. In such cases, going backward in time might require a different mathematical representation than going forward.

Analysts might react to backcasting the same way the White Queen did in Lewis Carroll's, *Through the Looking Glass*:

> "The White Queen lives backward through time. She begins to cry before she sticks herself with her brooch and stops immediately afterward. Living backward in time, she explains to Alice, 'always makes one a little giddy at first . . . But there's one great advantage to it - that one's memory works both ways.'

So you might be giddy about backcasting, but in the end, it is an empirical issue: Are the results from backcasting similar to those in forecasting? There have been few tests of backcasting. Armstrong (1985, p. 344) examined backcasts of sales of photographic goods for 17 countries by using data from 1965 to 1960 to forecast sales to 1954 (six years back); the accuracy for these backcasts was similar to that for forecasts to 1966 (six years ahead). Theil (1966, p. 177) compared an input-output model's accuracy for backcasts from one to eight years with its accuracy for forecasts of one to eight years; he obtained a close correspondence for two studies, one dealing with agriculture and the other with basic metal industries. Exhibit 2 shows that the root mean square errors for backcasts and forecasts were similar in Theil's study.

Exhibit 2
Backcast and Forecast Errors were Similar in Two Tests

Forecast Horizon (Years)	Root Mean Square Errors			
	Agriculture, Forestry, Fishing		Basic Metal Industries	
	Backcasts	Forecasts	Backcasts	Forecasts
1	4.0	4.0	10.6	10.4
2	6.0	6.2	15.7	15.4
3	7.9	7.9	20.1	21.5
4	9.9	9.5	20.7	25.7
6	12.0	11.6	29.9	32.2
8	12.5	11.0	39.8	48.2

- **Tailor analysis to the decision**

Tailor the analysis of forecasts to be useful to decision makers. Often this is an obvious step. However, when forecasting discrete events, or when asymmetries are involved, it may not be so clear.

Some analysts believe that error measures should account for asymmetries in the cost of errors. For example, forecast that are too low might lead to a loss of customers, which might have more serious consequences than forecasts that are too high. Leave this concern to the planners and decision makers. Forecasters should merely provide unbiased forecasts and good assessments of prediction intervals.

Sometimes it is possible to decompose the problem so decision makers can relate the errors to their problem. The late Allan Murphy, in a talk at the 1987 International Symposium on Forecasting, discussed the importance of decomposing the error. He used data provided by U.S. Army Sergeant J. P. Finley in 1884, where the task was to predict tornados. Exhibit 3 summarizes the results.

Exhibit 3
Value of Accuracy can Differ by its Use

	Actual Weather			
Forecasts	**Tornado**	**No Tornado**	**Totals**	**Hit Rate %**
Tornado	11	14	25	44.0
No Tornado	3	906	909	99.7
Totals	14	920	934	

Was Finley successful? He was correct for 917 of the 934 forecasts, which is 98.2%. But if he had always forecasted 'no tornado,' he would have been correct for 920 of the 934 cases, thus improving to 98.5%. People listening to the weather forecast, however, are most interested in what happens when tornados are forecast. Finley was successful in 44% (11 of 25) of the cases for which he forecasted tornados. He could have improved this score to 56% by never forecasting tornados, but decision makers would not have appreciated the improvement because the forecasts would have missed all 14 tornados that occurred. As it was, Finley correctly forecasted 11 of the 14.

TESTING THE DATA AND METHODS

Full disclosure is important in forecasting, as it is in any scientific research. In many areas of science, competing researchers sometimes withhold data and methodological details. Fortunately, academic researchers in forecasting normally disclose both.

- **Describe potential sources of bias by forecasters.**

Disclosure of conditions should include all important aspects of the study. One of the most important is the objectivity of the forecaster. For example, a researcher's proprietary interest in one of the methods should be disclosed, especially if judgment is involved in the forecast. Not surprisingly then, empirical studies in medicine are biased: drugs are found to be more effective when the study is done by someone with a proprietary interest (Armstrong 1997 provides a review). Brouthers (1986) found that errors in federal fiscal forecasts were biased by political party, ideology, and the year in the election cycle. Fildes and Hastings (1994), in their survey of nine divisions within a British multinational firm, found that 64% of the 45 respondents agreed that "forecasts are frequently politically modified."

Shamir (1986) classified 29 Israeli political surveys according to the independence of the pollster from low to high as "in-house," "commissioned," or "self-supporting;" the results showed that the more independent the pollster, the more accurate the predictions. Winston (1993) conducted a long-term follow-up of 30 published studies in which unbiased economists, using theory, made predictions about the effects of deregulation. Their predictions differed from those made by people affected by deregulation. The economists predicted that deregulation would be good for consumers, whereas those affected by the changes, who often were suspicious of the changes, predicted the opposite. As it turned out, the economists' unbiased predictions were much more accurate than the consumers'.

- **Assess the reliability and validity of the data.**

The reliability and validity of the input data limits one's ability to predict. These can affect the level or the forecast of change. For example, in 1986, the Joint Economic Committee of the U. S. Congress released a study showing an increase in the concentration of wealth. The committee estimated that the richest 0.5% of the families had held 25% of the wealth in 1963 and that, by 1983, the proportion had risen to 35%. The surveys were based on household samples, supplemented by samples of very rich persons selected by the Internal Revenue Service. Shortly

after the report was released, the finding for 1983 was found to be in error. One respondent, weighted heavily in the analysis, had been credited with $200 million of wealth whereas the correct figure was $2 million. When the figure was corrected, the estimated share of wealth held by the richest 0.5% dropped to 27% (Erickson 1988). Despite its prompt correction, the error caused many people to believe that wealth was too concentrated and that it was becoming more concentrated. Incidentally, the corresponding concentration figure in 1929 was 36%, so this longer time perspective might have led to the conclusion that wealth concentration was decreasing.

It is particularly important to test the reliability and validity of data used to assess policy changes. Card and Krueger (1994), using surveys of employment at New Jersey and Pennsylvania fast-food establishments, concluded that an increase in the minimum wage does not decrease employment among low-skilled workers. These economists were challenging an established principle in economics. Namely, that if the price of a good increases, the demand for that good will decrease. The study attracted much attention, probably because it reinforced what many people believe. Their findings became part of a State of the Union address by President Clinton. How could this study arrive at a conclusion that was contrary to the results from hundreds of studies on price elasticity, many of which related directly to minimum wages? As it turned out, Card and Krueger's data lacked reliability and validity, and the findings did not hold up when the study was replicated (Henderson 1996).

- **Provide easy access to data.**

Three arguments have been raised against making data freely available. First, the data might allow a competing researcher to receive credit for publications based on the data. Second, data can be costly to provide. Third, the researcher who has the data, especially company data, may wish to maintain confidentiality. For these reasons, obtaining data with which to replicate studies has sometimes been difficult in the management sciences.

With easy access to data, others can replicate an analysis. Sharing data advances science, and probably enhances the reputations of those who share. The M-competitions have been exemplary in their full disclosure. The data have been made available before and after the competitions. (The raw data for the M, M2, and M3-competitions are posted at the forecasting principles website, hops.wharton.upenn.edu/forecast.) The availability of these data has led to many replications and extensions. Partly as a result of this open procedure, the M-Competition paper by Makridakis et al. (1982) is the most widely cited paper on forecasting.

The argument that making data available is costly is less compelling now. Electronic storage is inexpensive and easy to access. However, confidentially remains a problem. Researchers can often preserve confidentiality by disguising and rescaling the data, although this may lead to a poorer description of the conditions. It makes a difference, for example, whether someone is forecasting weekly data on epidemics in China or annual sales of automobiles in Sweden.

- **Disclose details of the methods.**

When making generalizations about the use of various methods, it is important to describe the forecasting methods in detail. Those within the forecasting research community usually do this. The forecasting methods from the M-competitions have been made available, with two exceptions related to commercial entries.

When describing methods, authors should point out their deficiencies. This will aid others who study the problem. It should also help decision makers to determine how much confidence they can place in the findings. Weimann (1990) found that authors who described methodological shortcomings produced more accurate forecasts in political polls than did those who did not report shortcomings. Presumably, the researchers who published the shortcomings were more concerned about following proper methodology and thus did a better job.

- **Find out whether the clients understand the methods.**

Some consultants use complex methods. This practice, along with the use of needlessly complex language, "bafflegab," puts clients in an awkward position. Because the consultant spends time explaining the assumptions and apparently giving complete information about the model's structure, clients are reluctant to admit that they are so incompetent that they cannot understand the model. The easy way out is to nod in agreement and hope the consultant knows what he is talking about.

Consultants may use bafflegab to impress clients or to distract them from examining other stages of analysis. In our study of the FAITH models, we could not find a single person in the client's organization who

understood the models (Armstrong and Shapiro 1974). Typical comments were: “No one can explain FAITH to me” and “I don’t know how FAITH works.”

Consultants are not the only ones to use bafflegab. Academics have long used it as a way to gain prestige. Their readers are more impressed by obscure writing than by clear writing as shown in Armstrong (1980). In my opinion, obscure writing is becoming more pervasive and it can be seen in the forecasting journals.

Complexity is no virtue in forecasting and there is no excuse to use bafflegab when you have something to say. Consultants should explain their methods in ways that clients and potential users can understand. Clients should understand methods well enough to explain them to others. Consultants could use questionnaires or interviews to assess their clients’ understanding or perhaps ask clients to replicate the forecasts.

REPLICATING OUTPUTS

Despite the importance of replications, journals in the management sciences seldom publish them. Hubbard and Vetter (1996) analyzed 18 management journals from 1970 to 1991 and found few direct replications. Furthermore, extensions of prior studies typically make up less than ten percent of papers published in accounting, economics, and finance, and less than five percent of those in management and marketing. We have no evidence to suggest that the situation in forecasting is different, although the *International Journal of Forecasting* has a published policy to encourage replications.

- **Use direct replications to identify mistakes.**

In a direct replication, an independent researcher uses the same methods and the same data to determine whether they produce the same results. Direct replications may reveal mistakes because independent researchers are unlikely to make the *same* mistakes.

Armstrong and Shapiro (1974) reported on a direct replication. This used the same data and model that the FAITH consultants used. For 12 of the 15 periods, the forecasts by the FAITH consultants made were more accurate than those in the replication. Their average error was half that of the replication. The FAITH advocates were unable to explain this discrepancy, so the model failed this replication test.

Direct replications also help to ensure honesty, which should not be taken for granted in research. Cyril Burt, known as one of the world’s great psychologists, was famous for his study of the IQ scores of identical twins. Strangely, as Burt published accounts of his 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 cases. Wade (1976) describes this case. Some believed Burt was cheating, but I do not advise accusing a researcher of cheating. In my review of this issue (Armstrong 1983), it was difficult to find cases of cheating in academic studies in management science. Failures to replicate have many causes, and most, I suspect, are simply due to mistakes. Mistakes are common in academic research (Armstrong 1997).

Are mistakes common in forecasting? Gardner (1984) found 23 books and articles, mostly peer-reviewed, that had errors in model formulations for the trend in exponential smoothing. Errors may be more likely in work that is not subject to peer review. Gardner (1985) reported that mistakes were made in exponential smoothing programs used in two companies.

McLeavy, Lee and Adam (1981) replicated an earlier forecasting study by Adam. They found that two of the seven models in his original paper were in error. In this case, correcting the mistakes did not change the conclusions.

Adya (2000) conducted a direct replication of the Rule-Based Forecasting approach described by Collopy and Armstrong (1992). It revealed that six rules had been correctly applied but incorrectly reported, and four had been incorrectly applied. (The corrected rule-base is provided at the forecasting principles website.) Correction of these mistakes led to only small improvements in accuracy for the validation tests reported by Collopy and Armstrong.

- **Replicate studies to assess their reliability.**

If you apply the same methods to *similar data*, might you expect similar results? What if you apply *similar method*s to the same data? Either replication would provide evidence of reliability. Replications are much more useful as measures of reliability than are tests of statistical significance (Hubbard and Armstrong 1994).

The M-competition (Makridakis et al. 1982) was replicated by the M2-competition (Makridakis et al. 1993) and by the M3-competition (Makridakis and Hibon 2000). These used different sets of authors who tested different though similar time series for different time periods with minor variations in methods, criteria, and analyses. The major findings from the replications were consistent with those from the initial M-competition.

- **Extend studies to assess their generalizability.**

Extensions involve substantial changes from the original study such as using different data or conditions. For example, Fildes et al. (1998) used procedures from the M-competition and tested them on telecommunications data. The findings were similar.

You can generalize more easily if extensions produce similar findings. For example, the use of causal forces has been studied in different contexts with a variety of data sets. They have been shown to be useful for weighting extrapolation forecasts (Collopy and Armstrong 1992), selecting extrapolation methods (Armstrong & Collopy 1993), decomposing time series (Armstrong, Adya and Collopy 2001), and estimating prediction intervals (Armstrong & Collopy 2000).

Published extensions indicate that studies may not generalize to new situations. Hubbard and Vetter (1996) examined 266 replications and extensions in accounting, economics, finance, management, and marketing. The researchers who conducted the replications concluded that 27% of them provided full support for the original study and 27% partial support, but 46% of the findings were in conflict. While these results seem depressing, the situation in forecasting might be somewhat better. For example, researchers have performed many extensions of the M-competition study and these have supported the original findings.

- **Conduct extensions in realistic situations.**

The success of a method could be influenced by many factors, including the skills of the analysts, organizational politics, and the nature of the problem. It is useful, then, to examine whether the prescribed methods hold up well in field studies. Little research has been done on this topic. However, Bretschneider et al. (1989) surveyed various forecasting methods used by state governments. One of their conclusions was that agencies using simple econometric methods reported more accurate forecasts than those using more complex ones, thus reinforcing a conclusion from earlier studies.

Experimental field studies could be useful. For example, if a firm has twenty divisions, it might select ten on which to test a new forecasting procedure and use the other ten as a control group.

- **Compare forecasts obtained by different methods.**

Comparisons of forecasts from dissimilar methods or dissimilar data provide evidence on construct validity. When forecasts from different methods agree closely, the analyst gains confidence in them. If they differ substantially, the analyst should be suspicious about the forecasts.

ASSESSING OUTPUTS

- **Examine all important criteria.**

The selection of criteria represents a critical step in the evaluation of forecasting methods. Methods sometimes do well on one criterion but poorly on another, as Armstrong and Collopy (1992) showed in comparing the accuracy of extrapolation methods. Although accuracy is usually the primary concern, other criteria should also be considered. Especially important among these are timeliness, ease of interpretation, and ease of use. Forecasting experts, especially practitioners, regard ease of use as being nearly as important as accuracy (Yokum and Armstrong 1995).

- **Prespecify criteria.**

It is common for people, including researchers, to reject disconfirming evidence (e.g., Armstrong 1996; Batson 1975). They may misinterpret new evidence to confirm their forecast (Fischhoff 2001) or search until they find information that supports their forecast. They may even change criteria so as to support their forecast.

By specifying criteria before making a forecast, forecasters can avoid some of this behavior. It is especially important to prespecify criteria when forecasters have limited awareness of how they make predictions. In human resource decisions, for example, Webster (1964) found that people conducting job interviews typically make predictions rapidly, often in a few minutes, and with little awareness of what criteria and information they used.

To illustrate the importance of prior specification of criteria, we (Armstrong and Collopy 1994) examined the nomination of Clarence Thomas to the U.S. Supreme Court as it was in process. We expected that many people's decisions would be inconsistent with their criteria. The reconciliation of inconsistencies can be influenced by the way in which information is presented.

In mid-October 1991, shortly after the "weekend hearings," but before confirmation of the Thomas nomination, we asked 17 business school students if they would vote to confirm Thomas' nomination. All responses were obtained on self-administered anonymous questionnaires. They voted eight in favor, seven against, and two abstentions. We then posed two questions to them:

> Assume that you were designing a procedure to select a Supreme Court Justice. As part of the process you must make a selection between two candidates. Pool One contains 'Perjurers' (those who will lie under oath), while Pool Two contains 'Truth Tellers' (those who will only tell the truth under oath).
>
> Question 1: "What is the highest subjective probability that you would tolerate that your selection would come from Pool 1 (perjurers)?"
>
> Question 2: "What is your subjective probability that Clarence Thomas committed perjury?"

Their median acceptable level for selecting from the perjury pool was 13% and their median probability that Thomas was a perjurer was 45%. Then we asked the students to vote again. This time, they voted to reject, with six in favor to 11 against. Many stayed with their previous decision, even though it was inconsistent with their criteria.

If the problem can be structured so that the criteria are examined before predictions are made, people will be able to use the information more effectively. We expected that respondents would be more likely to judge Thomas a perjurer if they specified their criteria *before* they made their decisions about Clarence Thomas. So we asked a group of 46 business school students, "What is your subjective probability that Clarence Thomas committed perjury?" Their median estimate that Thomas was a perjurer was 92%. This estimate was considerably larger than the 45% estimate by subjects who had first been asked if they favored Thomas's appointment.

- **Assess face validity.**

In assessing face validity, ask experts whether forecasts look reasonable. You can also test the face validity of the model by using inputs that represent possible extremes that might be encountered. For example, in predicting automobile sales, consider what happens if you double income and halve the price of automobiles. If the forecasts look reasonable, this speaks well for the model. On the other hand, many forecasting models are designed to apply only within a certain range, so the failure to provide reasonable forecasts for extreme situations should not necessarily eliminate a method. Face validity can be a dangerous test because correct but unusual or unfavorable forecasts might be falsely judged to have low face validity.

Error Measures

The choice of an error measure can affect the ranking of methods. Exhibit 4 (from Armstrong & Collopy 1992) presents the agreement among accuracy rankings for 11 extrapolation methods. It was based on tests with 18 annual

time series selected by a probability sample from the M-competition data. The methods were ranked using six error measures. In general, the correlations among the rankings were not high, as their median was only .40. This means that the rankings of accuracy varied depending upon the choice of an error measure. For an extreme case, note that rankings by RMSE were negatively correlated to those by MdRAE at -.31.

Exhibit 4
Agreement Among Accuracy Rankings for 11 Methods is Sometimes Weak
(Spearman correlations for one-year horizon is for a set of 18 annual series)

Error Measure	**MAPE**	**MdAPE**	**Percent Better***	**GMRAE**	**Median RAE**
Root Mean Square Error (**RMSE**)	.44	.42	.11	.03	-.31
Mean Absolute Percentage Error (**MAPE**)		.83	.17	.68	.28
Median Absolute Percentage Error (**MdAPE**)			.09	.40	.06
Percent Better *				.46	.65
Geometric Mean of the RAE (**GMRAE**)					.79

* To keep the sign constant, we used "Percent Worse" rather than "Percent Better" for the correlation

- **Ensure that error measures are not affected by scale.**

When you compare alternative time-series methods, select error measures that are unaffected by scale. Otherwise, a few series with large numbers can dominate the comparisons, as they did in an analysis by Zellner (1986). He concluded that the Bayesian method was the most accurate in the M-competition comparisons because its RMSE was lowest. However, the RMSE is strongly influenced by scale, and Chatfield (1988), in a reexamination, showed that Zellner's conclusion resulted from large errors in only five of the 1,001 series.

Ensure error measures are valid.

Error measures should be valid for the task involved. They should make sense to experts (face validity) and should produce findings that agree with other measures of accuracy (construct validity). Armstrong and Collopy (1992) examined the agreement among various measures in accuracy rankings of 11 extrapolation methods based on 90 annual time series. The Spearman correlation between the RMSE and the consensus accuracy (based on average rankings of six accuracy measures) was .6. This was low compared to the rankings based on RAE or on MAPE, each of which had a correlation of .9 with the consensus. Once again, then, the RMSE performed poorly.

- **Avoid error measures with high sensitivity to the degree of difficulty.**

Some time series are harder to forecast than others. For example, it is easy to forecast the percentage of babies born in a hospital each month who will be boys because it remains fairly constant. On the other hand, it is difficult to forecast the daily sales of ice cream at a particular store in an area where the weather is changeable. To generalize about which forecasting methods are most appropriate for certain types of data, the error measure should not be highly affected by the degree of difficulty. Otherwise, too much weight may be placed on a few series that are difficult to forecast. This principle does not apply if the primary interest lies in forecasting the difficult cases, such as with wars, floods or hurricanes.

As a partial control for difficulty, one can correct for the amount of change occurring in a series. The assumption is that volatile series are more difficult to forecast. The Relative Absolute Error (RAE) is designed to control for change. It compares the error for a proposed forecasting model to that for the naive forecast.

For cross-sectional data, Ohlin and Duncan (1949) suggested the use of an index of predictive efficiency. You can create such an index by comparing the accuracy of a proposed model with an alternative, such as the current method's forecast, or the forecast using only the base rate.

- **Avoid biased error measures.**

When working with variables that contain only positive numbers, the MAPE is biased in that it favors low forecasts. By using an adjusted MAPE, one can correct this bias because the denominator is based not on the actual

outcome, but on the average of the actual outcome and the forecast. Armstrong (1985) describes characteristics of the adjusted MAPE.

Makridakis (1993) advocated the use of the adjusted MAPE (which he called the Unbiased Absolute Percentage Error or UAPE). He claimed that it would be more meaningful to managers, but this claim is speculative. The adjusted MAPE is less likely to be problematic when the actual value is zero or close to zero. Furthermore, it avoids the need for trimming. Given these advantages, it seemed that the adjusted MAPE might also be a reliable error measure. In an unpublished study, Fred Collopy and I conducted a small-scale test by comparing the rankings for subsets of data from annual series for the M-competition. We did this for five sets of 18 series each. We ranked 11 forecasting methods' accuracy for one-year-ahead and six-year-ahead forecasts, expecting the latter to be the most important test. There, the average rank correlation of the rankings with different data sets was .69 for the median adjusted APE and .72 for the median RAE (MdRAE). For one-year-ahead forecasts, the corresponding correlations were .75 for the median adjusted MAPE and .79 for the MdRAE. Thus, the adjusted MAPE offered no improvements in reliability in comparison with the MdRAE.

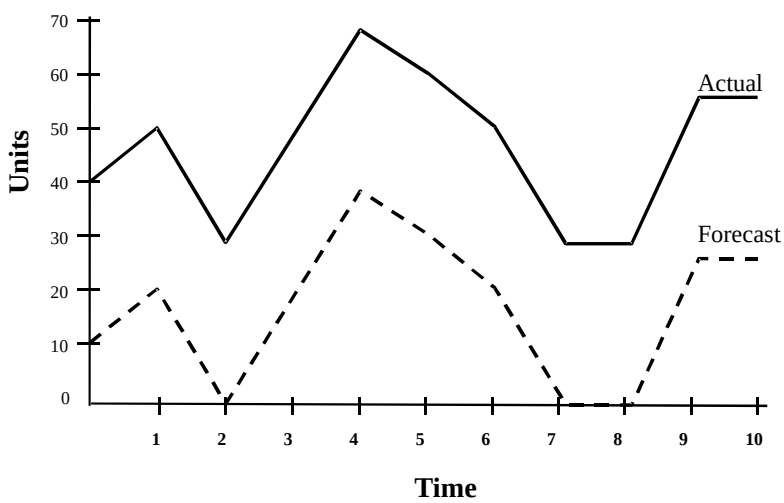
- **Avoid high sensitivity to outliers.**

When comparing alternative methods, avoid error measures that are highly sensitive to outliers. This suggests that the error measures should be trimmed, possibly even to the extent of using the median, which is the ultimate trimmed mean. This does not apply if the outliers are of primary interest. Nor does it apply when calibrating the parameters of models.

- **Do not use R-square to compare forecasting models**

R^2 should not be used for time series forecasts, not even on a forecast-validity sample. For one thing, it overlooks bias in forecasts. A model can have a perfect R^2, yet the values of the forecasts could be substantially different from the values for all forecasts, as shown in Exhibit 5. Also, a model could have an R^2 of zero but provide perfect forecasts if the mean were forecasted correctly and no variation occurred in the data. In addition, R^2 provides little information about effect size and thus has no obvious relationship to economic value.

Exhibit 5
High R^2 Does Not Ensure Accurate Forecasts



R^2 is misleading when used for time-series analysis. Ames and Reiter (1961) found high correlations for series that were randomly selected from the *Historical Statistics for the United States*. They regressed series against two to six randomly selected time-series variables. Using series of 25 years, they were usually able to "explain" over half of the variance.

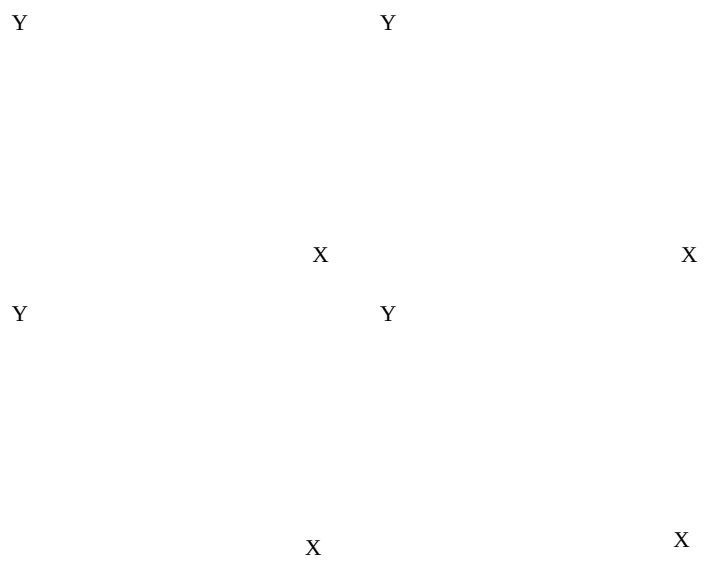
R^2 can also be misleading for cross-sectional forecasting. In one of my Tom Swift studies (Armstrong 1970), Tom analyzed data on sales of caribou chips among 31 countries. He used stepwise regression to select variables with t-statistics greater than 2.0, obtaining eight predictor variables from a possible total of 30. He dropped

three outliers from the analysis and obtained an adjusted R^2 of .85. That is interesting because all of the data had been drawn from a table of random numbers, so the true R^2 was zero.

I expect that the use of R^2 would mislead analysts in their comparison of forecasting models, although I have seen no studies on the issue. This follows from the work of Slovic and McPhillamy (1974). They concluded that when measures are easy to compare across alternatives, they receive too much emphasis, even when the selector believes that they are of little relevance. Fortunately, there are more useful measures of accuracy, so there is little need for R^2.

R^2 can also be misleading if used in the development of relationships for econometric models. Anscombe (1973), in an analysis of four alternative sets of data (sometimes referred to as Anscombe's quartet), showed that an equivalent R^2 (of .82)can be obtained from substantially different relationships.

Exhibit 6
R^2 May Not Provide a Good Description: Four Sets of Data with Equivalent R^2



Do not use Root Mean Square Error for comparisons across series.

Statisticians find the Root Mean Square Error (RMSE) attractive. Some analysts use it because large errors may have severe economic consequences. However, economic consequences should be evaluated by planners and decision makers, not by forecasters.

Research findings since the early 1980s have effectively ruled out the RMSE as a measure for comparing methods. Given that it is based on squaring forecast errors, it is unreliable, especially if the data might contain mistakes or outliers. In Armstrong and Collopy (1992), we ranked the accuracy of various methods when used to make forecasts for subsamples of annual and quarterly series from the M-competition (Makridakis et al. 1982). For example, 11 forecasting methods were ranked by accuracy when tested on 18 annual series. We then selected another 18 series and ranked the methods again for accuracy, calculating the correlation between the two sets of rankings. The process was repeated by selecting a third set of annual series, and making correlations between all three possible pairs. The process continued through five annual subsamples of 18 series each. Exhibit 7 presents the findings. For example, the average correlation for RMSE for eight-quarters ahead was -.13. Overall, the RMSE was unreliable, as indicated by its average correlation of .2. In contrast, Percent Better and the Relative Absolute Errors (MdRAE and GMRAE) provided rankings that were substantially more reliable, all being over .5.

Exhibit 7
Some Error Measures were Unreliable when Used to Rank 11 Forecasting Methods
(Average Spearman correlations for pairwise comparisons among five subsamples)

Error Measure	Quarterly*		Annual**		Average
	1-ahead	8-ahead	1-ahead	6-ahead	
RMSE	.14	–.13	.26	.54	.20
MdAPE	.13	.22	.46	.79	.40
MAPE	.59	.61	.49	.30	.50
GMRAE	.38	.17	.81	.74	.53
MdRAE	.39	.43	.79	.72	.58
Percent Better	.60	.59	.82	.78	.70

* 20 series per subsample ** 18 series per subsample

Pant and Starbuck (1990) provided evidence on the validity of RMSE. They compared the fit and forecast errors for series from the M-Competition. For one-period-ahead forecasts, they found a negative rank correlation (r = -.11) between fit and forecast errors when they examined 21 methods. When they used a six-period-ahead horizon, the correlation was again negative (–.08). When they restricted their analysis to the 13 methods that used seasonal adjustments, a more reasonable test, the correlations were still negative (–.15 for one-ahead and –.20 for six-ahead). They standardized the data and again tested the 13 seasonal methods. This provided little help as the one-period-ahead correlation was +.16 and the six-ahead was -.18.

Given its poor reliability and validity, the RMSE should not be used for comparisons even in large-sample studies, such as the 1,001 series used in the M-Competition. Armstrong and Fildes (1995) summarized evidence related to the use of the RMSE.

Summarizing and rating the error measures: Interestingly, while statisticians seem to pay little attention to empirical studies on error measures (Armstrong and Fildes 1995), forecasters have apparently changed their beliefs. In a 1981 survey by Carbone and Armstrong (1982), the RMSE was preferred by 48% of 62 academics and 33% of 61 practitioners. This made RMSE the most popular of the seven error measures listed. At the same time, the MAPE was used by 24% of academics and only 11% of practitioners. In contrast, over a decade later, in a survey of 207 forecasting executives, Mentzer and Kahn (1995) found that MAPE was the most commonly used measure (52%) while only 10% used RMSE.

Exhibit 8 summarizes error measures for time series, along with ratings on some criteria. (The measures are defined in the appendix to this paper.) This list of measures is not comprehensive. For example, some experts advocate the use of turning-point errors. However, using turning-point errors in comparing methods typically yields small samples and it provides no information on the magnitude of errors.

Exhibit 8
Ratings of Error Measures

Error measure	Reliability	Construct validity	Outlier protection	Control for difficulty?
Root Mean Square Error	poor	fair	poor	no
Percent Better	good	fair	good	yes
Mean Absolute Percentage Error	fair	good	poor	no
Median Absolute Percentage Error	fair	good	good	no
Geometric Mean Relative Absolute Error	fair	good	fair	yes
Median Relative Absolute Error	fair	good	good	yes

- **Use multiple error measures.**

Some error measures are more useful than others. However, after examining the situation, you may find more than one error measure that is relevant. In such a case, decide in advance which is the most relevant and focus on it, but also present results for other measures that might be relevant. The M-competition studies have, for example, routinely provided results based on a variety of error measures.

- **Use ex ante tests of forecast accuracy.**

Milton Friedman (see Friedman & Schwartz 1991) reported on his development, in 1945, of an econometric model. He had been asked to analyze data on alloys used in turbine blades for engines. The goal was to develop an alloy that would withstand high temperatures for long periods. He used standard regression analysis to develop an econometric model that predicted time to failure as a function of stress, temperature, and variables representing the alloy's composition. Obtaining estimates for this equation, along with associated statistics, would have taken a highly skilled analyst about three months. Fortunately, however, there was one large computer in the country – the Mark 1 at Harvard – that could do the calculations. This computer, built from many IBM card-sorting machines, was housed in an air-conditioned gymnasium. Ignoring the time for data input, it required 40 hours to calculate the regression. A regression of this size could be solved in less than a second on a desktop computer today. Friedman was delighted with the results. The model had a high R^2, and it performed well on all relevant statistics. The model led him to recommend two new improved alloys. It predicted that each alloy would survive several hundred hours at high temperatures. Tests of the new alloys were carried out in a lab at MIT and each ruptured in less than four hours. Friedman concluded that statistical measures of a model's ability to fit historical data provide little evidence about its ability to predict with new data.

Friedman's experience does not imply that measures of fit are never useful. Certainly they are of some help. For example, a model that cannot explain any historical variation is unlikely to be useful in forecasting. Furthermore, measures of fit for cross-sectional data have been useful in some areas, such as in personnel selection (Barrett, Phillips and Alexander 1981).

A long stream of research shows that the fit of a model to time-series data (the calibration sample) provides a poor way to assess predictive validity. Ferber (1956) examined forecasts of the total savings in the U.S. He calibrated seven models using data from 1923-1940 and made forecasts for 1947-1949; there was only a small relationship between the R^2 of the calibration sample and the forecast accuracy. Schupack (1962), in a study of short-range forecasts for food and household products, found only a slight relationship between the fits of regression models and their accuracy. Mayer (1975) examined relationships between fit and forecast errors for 99 comparisons published in economics journals from 1954 to 1975; he concluded that the calibration fit was a poor guide to forecast accuracy. Elliott and Baier (1979) obtained excellent explanations for changes in interest rates, with R^2 of the fits as high as .996; however, for one-month-ahead forecasts, their models were inferior to the naive (no-change) model. Sexton (1987) compared the accuracy of six econometric and extrapolation models in forecasting changes in the value of residential real estate in 77 Minnesota counties. He used data from 1974 through 1980 to fit the models and made ex post forecasts for one, two, and three-year horizons; the fit to historical data had little relationship to forecast accuracy in the models.

Pant and Starbuck (1990) found a modest relationship between fit and forecast accuracy. They did this by comparing 13 seasonal models in making predictions for the 1,001 time series of the M-competition. Using MAPE, they found that rankings based on fit correlated + .49 with ratings based on one-period-ahead ex ante forecast accuracy. For six-period-ahead forecasts, the corresponding correlation was +.2.

Perhaps the most reasonable position, given the evidence to date, is that the calibration fit for time-series models provides a weak measure of forecast validity. Instead of fit, you should try to closely simulate the actual forecasting situation and then compare the methods using ex ante forecast errors.

With ex ante tests, methods are used to make forecasts without any knowledge of what happened in the actual situation. For extrapolation methods, this is normally accomplished by withholding the most recent data; even though the events have occurred, the researcher can assume that they have not. This seems to be a useful assumption. When it was used in the M-Competition, however, some critics objected on the basis that the events had already occurred. To address this issue, Makridakis conducted the M2-Competition in real time, that is, before the events actually occurred. The findings from the M2-Competition were similar to those from the M-Competition (Makridakis et al. 1993). This supports the procedure of withholding data to assess the forecast validity of extrapolation methods.

Tests of judgmental forecasting are more likely to be compromised by knowledge of the outcomes. It may be difficult for experts to make forecasts about things that have already occurred without having their knowledge of outcomes affect their forecasts. It is also difficult to get people to accurately recall what forecasts they made in the past (Fischhoff 2001 reviews research on hindsight bias). Sometimes, however, researchers can disguise historical situations, as I did in role-playing studies (Armstrong 2001a).

- **Use statistical significance only to compare the accuracy of reasonable forecasting models.**

According to Cohen's (1994) review of studies in psychology, researchers often misinterpret statistical significance. McCloskey and Ziliak (1996) reached similar conclusions in their examination of papers published in leading economics journals. One problem is that researchers (and editors and reviewers) often confuse statistical significance with practical significance. As a result, some leading researchers have recommended that journals ban the use of significance tests in published studies. To my knowledge, only one journal did so, and the results seemed to be beneficial (Armstrong 1997).

Statistical significance may be useful when using small validation samples to judge whether accuracy differs among reasonable forecasting methods. Here, the null hypothesis that the methods are equally accurate is often a reasonable alternative. For large samples, such as the M-competition, nearly all differences among methods would be statistically significant, so it offers little help in discriminating among them.

If you use significance tests, adjust them for the number of methods that are being compared. To do this, use the tables provided by Dunnett (1955, 1964) (reproduced in Armstrong 1985, pp. 462-467 and provided on the forecasting principles website). The more methods you compare, the more likely it is that traditional tests will show one of them to be significantly better than the others. Power tests might also be useful (Cohen 1988) for assessing what sample size you will need to identify large, medium, or small differences.

- **Use ex post comparisons to test policy forecasts.**

Ex post comparisons are important for assessing how well one can predict the effects of policy changes. That is, if one knows the true state of the world in the future, to what extent would forecasts of policy changes prove to be accurate? Suppose that a marketing plan called for a 15% reduction in the price of a product, and this was incorporated in the forecast. Now assume that, early in the forecast horizon, management decided to cancel the price reduction. You might say that the forecaster should have forecasted management's behavior. However, if you want to predict the effects of a price change in the future, you should analyze ex post forecasts. That is, forecasters should use actual changes in the causal variables, then compare forecasts and actual values.

- **Obtain a large sample of independent forecast errors.**

Sample sizes are sometimes inadequate. For example, Focus Forecasting is deficient in that its tests use only three monthly forecasts (Gardner and Anderson 1997). This leads to an unreliable procedure for testing alternative methods. Does this matter? Two published studies conducted evaluations of Focus Forecasting. Flores and Whybark (1986) used 96 actual time series and found Focus Forecasting to be substantially less accurate than exponential smoothing. Gardner (1997) compared Focus Forecasting to damped trend exponential smoothing for 68 monthly and 23 quarterly series drawn from the M-competition, and five monthly cookware series from a Houston firm. Focus Forecasting proved substantially less accurate than damped smoothing in these comparisons. For the M-competition data, its Median APE was 18% higher for the monthly series, and 32% higher for the quarterly series. It was 44% higher for the monthly cookware series.

In comparing methods, one prefers a large sample of independent forecasts. The best way to obtain independent time-series forecasts is to use time series that differ substantially. How many series are needed? Armstrong and Collopy (1992) examined reliability using samples of 18 annual series and concluded that reliability was a problem with samples this small. Makridakis and Hibon (1979) pioneered the use of large sample comparisons when they published their study of 111 time series. Some people claimed that this number of series was too small, so Makridakis organized larger studies, including the M-competition data set, which has 1,001 series (Makridakis et al. 1982) and the M3-competition which contains 3,003 time-series (Makridakis and Hibon 2000).

There are other ways to increase sample sizes. Given a single origin, one can forecast for a number of horizons. For example, in the M-Competition, forecasts were made for each of six years ahead (as well as up to eight

quarters ahead and 18 months ahead) for each of the series. The forecast errors for different horizons are expected to be related to one another, especially for short-interval data, such as monthly data. An alternative way to examine these errors is to look at the cumulative sum of the absolute errors over the horizon, the Cumulative Relative Absolute Error (CumRAE), which is described in the appendix of this paper.

Another approach to increase the number of forecasts is to use successive updating, sometimes called a rolling horizon (or moving origin). Forecasts are made from the origin *t*, say 1990, for the next six years. The next year, 1991, is then included in the calibration. After updating, forecasts are made for the next six years. The procedure (Exhibit 9) continues until one has no more data left in the forecast validation sample. This procedure does not provide independent observations, especially for short-intervals such as monthly data. The results should be used with caution, and statistical significance is of little value. Successive updating has been employed in published studies at least as far back as the early 1970s.

Exhibit 9
Successive Updating (Rolling Horizon)

Calibration Data Ends With:	**Forecast Periods**					
	***t* + 1**	***t* + 2**	***t* + 3**	***t* + 4**	**. . .**	***t* + *h***
t	F_1	F_2	F_3	F_4	. . .	F_h
$t + 1$	–	F_1	F_2	F_3	. . .	F_h-1
$t + 2$	–	–	F_1	F_2	. . .	F_h-2
. . .	. . .	. . .	. . .	. . .	. . .	. . .
$t + h - 1$	–	–	–	–	. . .	F_1

For cross-sectional data, one validation procedure is to split the data into two parts. Forecasting models are developed on the calibration set and tested on the other part. However, researchers can increase sample sizes by using a jackknife procedure. With this procedure, the researcher uses all but one of the observations to develop the model and makes a forecast for the excluded observation. The observation is then included, while another is excluded. The researcher reestimates the model on the new calibration sample and makes a forecast for the excluded observation. The procedure is repeated until forecasts have been made for each observation. Thus, one can obtain 100 holdout forecasts from a sample of 100 observations.

While sample size is important, its value can be overstated, as sampling error is only part of the total forecast error. Nonresponse bias and response errors can be substantial. For example, in Lau's (1994) study of 56 political polls, sample size varied from 575 to 2,086. Despite this range, sample size had little relationship to the eventual accuracy of the poll. Its correlation to forecast accuracy was small, only .1 in this sample. However, sampling error would be important when small samples are used. Consider, for example, the enormous errors if one were to generalize from the results of three focus groups.

- **Conduct explicit cost-benefit analyses.**

Given a set of applicable models with realistic assumptions, the general procedure for a cost-benefit analysis is to list the potential benefits and costs for each model, find some scheme to compare them, and then rank them by a cost/benefit score.

The costs include initial development costs, maintenance costs (to keep the model up-to-date), and operating costs (time and dollars to make the forecasts). The benefits include improved forecast accuracy, better assessments of uncertainty, and consistent evaluations of alternative futures (including changes in the environment or changes in the organization's policies). Unfortunately, it is not easy to estimate the value of each benefit. It is even difficult to forecast the value of improved accuracy, although a framework for such an evaluation is provided in the practitioners' section of the forecasting principles website.

IMPLICATIONS FOR PRACTITIONERS

Informal trials are unlikely to lead to proper conclusions about which forecasting method is best. In such trials, people tend to confirm their existing beliefs. Evaluating alternative forecasting methods is a complex task. I suggest the systematic use of a checklist to ensure application of evaluation principles (Exhibit 10).

Exhibit 10
Evaluation Principles Checklist

Using reasonable alternatives
- Compare reasonable forecasting methods

Testing assumptions
- Use objective tests of assumptions
- Test assumptions for construct validity
- Describe conditions of the problem
- Match tests to the problem
- Tailor analysis to the decision

Testing data and methods
- Describe potential biases
- Assess reliability and validity of data
- Provide easy access to data
- Disclose details of methods
- Find out whether clients understand the methods

Replicating outputs
- Use direct replication to identify mistakes
- Replicate studies to assess reliability
- Extend studies to assess generalizability
- Conduct extensions in realistic situations
- Compare with forecasts obtained by different methods

Assessing outputs
- Examine all important criteria
- Prespecify criteria
- Assess face validity
- Adjust error measures for scale
- Ensure error measures are valid
- Avoid error measures sensitive to degree of difficulty
- Avoid biased error measures
- Avoid sensitivity to outliers
- Do not use R^2 to compare models
- Do not use Root Mean Square Error
- Use multiple error measures
- Use ex ante tests for accuracy
- Use statistical significance only to test accuracy of reasonable models
- Use ex post tests for policy effects
- Obtain large samples of independent forecast errors
- Conduct explicit cost/benefit analysis

IMPLICATIONS FOR RESEARCHERS

Most of the principles for evaluating forecasting methods have been drawn from standard procedures in the social sciences, and many of these are based on common sense or received wisdom. As a result, there is little need for research on the latter ones. That applies to the principles in the left-hand column of Exhibit 10, except perhaps for identifying whether tests of construct validity of data relate to performance, how the reliability and validity of data relate to accuracy, whether tests of construct validity are important, and the extent to which client's understanding of methods affects their use. On the other hand, the principles in the right column would all benefit from further research, except perhaps for the principles related to R^2 and root mean square error, as these have already been subjected to much study, and for explicit cost/benefit analysis, which is based on common sense.

CONCLUSIONS

Principles for evaluation can help forecasters select methods for a given situation. They can also be useful to those conducting academic research on forecasting. Most of the principles are based on standard research methodology, and empirical evidence exists for only some of them. Despite the intuitive appeal of these principles, practitioners and academic researchers often ignore many of them. Neglected principles include obtaining a large sample of independent forecasts, describing conditions, and conducting replications. Some evaluation principles are counterintuitive. In particular, there is a poor correspondence between statistical fit and forecast accuracy for time-series data.

REFERENCES

Adya, M. (2000), "Corrections to rule-based forecasting: Results of a replication," *International Journal of Forecasting*, 16, 125-127.

Ames, E. & S. Reiter (1961), "Distributions of correlation coefficients in economic time series, " *Journal of the American Statistical Association*, 56, 637-656.

Anscombe, F. J. (1973), "Graphs in statistical analysis," *American Statistician*, 27, 17-21.

Armstrong, J. S. (1970), "How to avoid exploratory research," *Journal of Advertising Research*, 10 (August), 27-30. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1979), "Advocacy and objectivity in science," *Management Science*, 25, 423-428.

Armstrong, J. S. (1980), "Unintelligible management research and academic prestige," *Interfaces*, 10 (March-April), 80-86. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1983), "Cheating in management science," *Interfaces*, 13 (August), 20-29.

Armstrong, J. S. (1984), "Forecasting by extrapolation: Conclusions from 25 years of research," *Interfaces*, 13 (Nov./Dec.), 52-61. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1985), *Long-Range Forecasting*. New York: John Wiley. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1988), "Research needs in forecasting," *International Journal of Forecasting*, 4, 449-465. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. (1996), "Management folklore and management science: On portfolio planning, escalation bias, and such," *Interfaces*, 26, No. 4, 28-42.

Armstrong, J. S. (1997), "Peer review for journals: Evidence on quality control, fairness, and innovation," *Science and Engineering Ethics*, 3, 63-84. Full text at hops.wharton.upenn.edu/forecast. See "peer review."

Armstrong, J. S. (2001a), "Role-playing: A method to forecast decisions," in J. S. Armstrong (ed.), *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S. (2001b), "Selecting forecasting methods," in J. S. Armstrong (ed.), *Principles of Forecasting.* Norwell, MA.: Kluwer Academic Publishers.

Armstrong, J. S., M. Adya & F. Collopy (2001), "Rule-based forecasting: Using judgment in time-series extrapolation," in J. S. Armstrong (ed.), *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S., R. Brodie & A. Parsons (2001), "Hypotheses in marketing: Literature review and publication audit," *Marketing Letters* (forthcoming).

Armstrong, J. S. & F. Collopy (1992), "Error measures for generalizing about forecasting methods: Empirical comparisons," *International Journal of Forecasting*, 8, 69-80. Full text at hops.wharton.upenn.edu/forecast.

Followed by commentary by Ahlburg, Chatfield, Taylor, Thompson, Winkler and Murphy, Collopy and Armstrong, and Fildes, pp. 99-111.

Armstrong, J. S. & F. Collopy (1993), "Causal forces: Structuring knowledge for time series extrapolation," *Journal of Forecasting*, 12, 103-115. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. & F. Collopy (1994), "How serious are methodological issues on surveys? A reexamination of the Clarence Thomas polls." Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. & F. Collopy (2000), "Identification of asymmetric prediction intervals through causal forces" *Journal of Forecasting* (forthcoming).

Armstrong, J. S. & R. Fildes (1995), "On the selection of error measures for comparisons among forecasting methods," *Journal of Forecasting*, 14, 67-71. Full text at hops.wharton.upenn.edu/forecast.

Armstrong, J. S. & A. Shapiro (1974), "Analyzing quantitative models," *Journal of Marketing*, 38, 61-66. Full text at hops.wharton.upenn.edu/forecast.

Barrett, G. V., J. S. Phillips & R. A. Alexander (1981), "Concurrent and predictive validity designs: A critical reanalysis," *Journal of Applied Psychology*, 66, 1-6.

Batson C. D. (1975), "Rational processing or rationalization? The effect of disconfirming information on a stated religious belief," *Journal of Personality and Social Psychology*, 32, 176-184.

Bretschneider, S. I., W. L. Gorr, G. Grizzle & E. Klay (1989), "Political and organizational influences on the accuracy of forecasting state government revenues," *International Journal of Forecasting*, 5, 307-319.

Brouthers, L. E. (1986), "Parties, ideology, and elections: The politics of federal revenues and expenditures forecasting," *International Journal of Public Administration*, 8 (3), 289-314.

Brown, R. G. (1959), *Statistical Forecasting for Inventory Control*. New York: McGraw-Hill.

Carbone, R. & J. S. Armstrong (1982), "Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners," *Journal of Forecasting*, 1, 215-217. Full text at hops.wharton.upenn.edu/forecast.

Card, D. & A.B. Krueger (1994), "Minimum wages and a case study of the fast-food industry in New Jersey and Pennsylvania," *American Economic Review*, 84, 772-793.

Chamberlin, C. (1965), "The method of multiple working hypotheses," *Science*, 148, 754-759.

Chatfield, C. (1988), "Apples, oranges and mean square error," *Journal of Forecasting*, 4, 515-518.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Cohen, J. (1994), "The earth is round ($p < .05$)," *American Psychologist*, 49, 997-1003.

Collopy, F., M. Adya & J. S. Armstrong (1994), "Principles for examining predictive validity: The case of information systems spending forecasts," *Information Systems Research*, 5, 170-179.

Collopy, F. & J. S. Armstrong (1992), "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations," *Management Science*, 38, 1394-1414.

Dalessio, A. T. (1994), "Predicting insurance agent turnover using a video-based situational judgment test," *Journal of Business and Psychology*, 9, 23-37.

Dunnett, C. W. (1964), "New tables for multiple comparisons with a control," *Biometrics*, 20, 482-491.

Dunnett, C. W. (1955), "A multiple comparison procedure for comparing several treatments with a control," *Journal of the American Statistical Association*, 50, 1096-1121.

Elliott, J. W. & J. R. Baier (1979), "Econometric models and current interest rates: How well do they predict future rates?" *Journal of Finance*, 34, 975-986.

Erickson, E.P. (1988), "Estimating the concentration of wealth in America," *Public Opinion Quarterly*, 2, 243-253.

Ferber, R. (1956), "Are correlations any guide to predictive value?" *Applied Statistics*, 5, 113-122.

Fildes, R. (1992), "The evaluation of extrapolative forecasting methods," *International Journal of Forecasting*, 8, 81-98.

Fildes, R. & R. Hastings (1994), "The organization and improvement of market forecasting," *Journal of the Operational Research Society*, 45, 1-16.

Fildes, R., M. Hibon, S. Makridakis & N. Meade (1998), "Generalizing about univariate forecasting methods: Further empirical evidence" (with commentary), *International Journal of Forecasting*, 14, 339-366.

Fildes, R. & S. Makridakis (1988), "Forecasting and loss functions," *International Journal of Forecasting*," 4, 545-550.

Fischhoff, B. (2001), "Learning from experience: Coping with hindsight bias and ambiguity," in J. S. Armstrong (ed.), *Principles of Forecasting*. Norwell, MA: Kluwer Academic Publishers.

Flores, B. & C. Whybark (1986), "A comparison of focus forecasting with averaging and exponential smoothing," *Production and Inventory Management*, 27, (3), 961-103.

Friedman, M. (1953), "The methodology of positive economics," *Essays in Positive Economics.* Chicago: University of Chicago Press.

Friedman, M. & A. J. Schwartz (1991), "Alternative approaches to analyzing economic data." *American Economic Review*, 81, Appendix, pp. 48-49.

Gardner, E. S., Jr. (1984), "The strange case of lagging forecasts," *Interfaces*, 14 (May-June), 47-50.

Gardner, E. S., Jr. (1985), "Further notes on lagging forecasts," *Interfaces*, 15 (Sept-Oct.), 63.

Gardner, E. S., Jr. & E. A. Anderson (1997), "Focus forecasting reconsidered," *International Journal of Forecasting*, 13, 501-508.

Gurbaxani, V. & H. Mendelson (1990), "An integrative model of information systems spending growth," *Information Systems Research*, 1, 254-259.

Gurbaxani, V. & H. Mendelson (1994), "Modeling vs. forecasting: The case of information systems spending," *Information Systems Research*, 5, 180-190.

Henderson, D.R. (1996), "Rush to judgment," *Managerial and Decision Economics*, 17, 339-344.

Hubbard, R. & J. S. Armstrong (1994), "Replications and extensions in marketing: Rarely published but quite contrary," *International Journal of Research in Marketing*, 11, 233-248. Full text at hops.wharton.upenn.edu/forecast.

Hubbard, R. & D. E. Vetter (1996), "An empirical comparison of published replication research in accounting, economics, finance, management, and marketing," *Journal of Business Research*, 35, 153-164.

Lau, R. D. (1994), “An analysis of the accuracy of ‘trial heat’ polls during the 1992 presidential election,” *Public Opinion Quarterly*, 58, 2-20.

Machlup, F. (1955), “The problem of verification in economics,” *Southern Economic Journal*, 22, 1-21.

Makridakis, S. (1993), “Accuracy measures: Theoretical and practical concerns,” *International Journal of Forecasting*, 9, 527-529.

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord & L. F. Simmons (1993), “The M2-competition: A real-time judgmentally based forecasting study,” *International Journal of Forecasting*, 9, 5-22 (commentary follows on pages 23-29).

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen & R. Winkler (1982), “The accuracy of extrapolation (time series) methods: Results of a forecasting competition,” *Journal of Forecasting*, 1:111-153.

Makridakis, S. & M. Hibon (2000), “The M3-competition: Results, conclusions and implications,” *International Journal of Forecasting*, (forthcoming).

Makridakis, S. & M. Hibon (1979), “Accuracy in forecasting: An empirical investigation” (with discussion), *Journal of the Royal Statistical Society: Series A*, 142, 97-145.

Mayer, T. (1975), “Selecting economic hypotheses by goodness of fit,” *The Economic Journal*, 85, 877-883.

McCloskey, D. N. & S. T Ziliak, (1996), “The standard error of regressions,” *Journal of Economic Literature*, 34, 97-114.

McLeavy, D.W., T.S. Lee & E. E. Adam, Jr. (1981), “An empirical evaluation of individual item forecasting models” *Decision Sciences*, 12, 708-714.

Meade, N. (1995), “Research on forecasting,” *International Journal of Forecasting*, 11, 355-358.

Mentzer, J. T. & K. B. Kahn (1995), "Forecasting technique familiarity, satisfaction, usage, and application,” *Journal of Forecasting*, 14, 465-476.

Nagel, E. (1963), “Assumptions in economic theory,” *American Economic Review*, 53, 211-219.

Ohlin, L. E. & O. D. Duncan (1949), “The efficiency of prediction in criminology,” *American Journal of Sociology*, 54, 441-452.

Pant, P. N. & W. H. Starbuck (1990), “Innocents in the forest: Forecasting and research methods,” *Journal of Management*, 16, 433-460.

Schnaars, S. (1984), “Situational factors affecting forecast accuracy,” *Journal of Marketing Research*, 21, 290-297.

Schupack, M. R. (1962), “The predictive accuracy of empirical demand analysis,” *Economic Journal*, 72, 550-575.

Sexton, T. A. (1987), “Forecasting property taxes: A comparison and evaluation of methods,” *National Tax Journal*, 15, 47-59

Shamir, J. (1986), “Pre-election polls in Israel: Structural constraints on accuracy,” *Public Opinion Quarterly*, 50, 62-75.

Slovic, P. & D. J. McPhillamy (1974), “Dimensional commensurability and cue utilization in comparative judgment,” *Organizational Behavior and Human Performance*, 11, 172-194.

Smith, B. T. (1978), *Focus Forecasting: Computer Techniques for Inventory Control*. Boston: CBI Publishing.

Smith, M. C. (1976), "A comparison of the value of trainability assessments and other tests for predicting the practical performance of dental students," *International Review of Applied Psychology*, 25, 125-130.

Stephan, W. G. (1978), "School desegregation: An evaluation of predictions made in Brown v. Board of Education," *Psychological Bulletin*, 85, 217-238.

Theil, H. (1966), *Applied Economic Forecasting*. Chicago: Rand McNally.

Wade, N. (1976), "IQ and heredity: Suspicion of fraud beclouds classic experiment," *Science*, 194, 916-919.

Webster, E. C. (1964), *Decision Making in the Employment Interview*. Montreal: Eagle.

Weimann, G. (1990), "The obsession to forecast: Pre-election polls in the Israeli press," *Public Opinion Quarterly*, 54, 396-408.

Winston, C. (1993), "Economic deregulation: Days of reckoning for microeconomists," *Journal of Economic Literature*, 31, 1263-1289.

Yokum, T. & J. S. Armstrong (1995), "Beyond accuracy: Comparison of criteria used to select forecasting methods," *International Journal of Forecasting*, 11, 591-597. Full text at hops.wharton.upenn/edu/forecast.

Zellner, A. (1986), "A Tale of forecasting 1001 series: The Bayesian knight strikes again," *International Journal of Forecasting*, 2, 491-494.

Acknowledgments: P. Geoffrey Allen, Fred Collopy, Don Esslemont, Peter S. Fader, Richard H. Franke, Raymond Hubbard, Donald G. MacGregor, Nigel Meade, Herman O. Stekler, Leonard J. Tashman and J. Thomas Yokum provided helpful suggestions on early drafts. Raphael Austin, Ling Qiu and Mariam Rafi provided editorial assistance.

May 19, 2000

APPENDIX

The following notation is used for the definitions of error measures that follow:

m	is the forecasting method,
rw	is the random walk method,
h	is the horizon being forecast,
s	is the series being forecast,
$F_{m,h,s}$	is the forecast from method m for horizon h of series s,
$A_{h,s}$	is the actual value at horizon h of series s,
H	is the number of horizons to be forecast, and
S	is the number of series being summarized.

The absolute percentage error (APE) for a particular forecasting method for a given horizon of a particular series is defined as

$$APE_{m,h,s} = \left| \frac{F_{m,h,s} - A_{h,s}}{A_{h,s}} \right|$$

The APEs for a particular forecasting method are summarized across series by

$$MAPE \frac{\sum_{s-1}^{S} APE_{m,h,s}}{S} x\,100 \text{ or by}$$

$MdAPE_{m,h}$ = Observation $\frac{S+1}{2}$ if S is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by $APE_{m,h,s.}$

The relative absolute error (RAE) for a particular forecasting method for a given horizon of a particular series is defined as

$$RAE_{m,h,s} = \frac{\left| F_{m,h,s} - A_{h,s} \right|}{\left| F_{rw,h,s} - A_{h,s} \right|}$$

The Winsorized RAEs are defined by

$$WRAE_{m,h,s} = \begin{cases} 0.01 \text{ if } RAE_{m,h,s} < 0.01 \\ RAE_{m,h,s} \text{ if } 0.01 \le RAE_{m,h,s} \le 10 \\ 10 \text{ if } RAE_{m,h,s} > 10 \end{cases}$$

Because we always recommend Winsorizing of the RAE, we drop the *W* below and in the text. The Winsorized RAEs for a particular forecasting method are summarized across series by

$$GMRAE_{m,h} = \left[\prod_{s-1}^{S} RAE_{m,h,s} \right]^{1/S} \text{ or by}$$

$MdRAE_{m,h} =$ Observation $\frac{S+1}{2}$ if *S* is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by $RAE_{m,h,s}$

The RAEs for a particular forecasting method are summarized across all of the *H* horizons on a particular series by

$$CumRAE_{m,s} = \frac{\sum_{h=1}^{H} \left| F_{m,h,s} - A_{h,s} \right|}{\sum_{h=1}^{H} \left| F_{rw,h,s} - A_{h,s} \right|}$$

The CumRAE is Winsorized in the same way as the GMRAE is Winsorized. The CumRAEs for a particular forecasting method are summarized across series by

$$GMCumRAE_{m} = \left[\prod_{s=1}^{S} CumRAE_{m,s} \right]^{1/S} \textit{ or by}$$

$MdCumRAE_{m} =$ Observation $\frac{S+1}{2}$ if *S* is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by Winsorized $CumRAE_{m,s}$

The root mean squared errors (RMSEs) for a particular forecasting method are summarized across series by

$$RMSE_{m,h} = \left(\frac{\sum_{s=1}^{S} \left(F_{m,h,s} - A_{h,s} \right)^2}{S} \right)^{1/2}$$

Percent Better is calculated as

$$\text{Percent Better}_{\text{m,h}} = \frac{\sum_{s=1}^{S} j_s}{S} x100$$

$$\text{where } j_s = \begin{matrix} 1 \text{ if } \left| F_{m,h,s} - A_{h,s} \right| < \left| F_{rw,h,s} - A_{h,s} \right| \\ 0 \text{ otherwise.} \end{matrix}$$