Published in *International Journal of Forecasting*, 17, 2001, 542-545.

Should We Redesign Forecasting Competitions?

J. Scott Armstrong
The Wharton School, University of Pennsylvania

The M3-Competition continues to improve the design of forecasting competitions: It examines more series than any previous competition, improves error analyses. and includes commercial forecasting programs as competitors. To judge where to go from here, I step back to look at the M-Competitions as a whole. I discuss the advantages of the M-Competitions in hopes that they will be retained, describe how to gain additional benefit from future competitions, and finally, describe a low-cost approach to competitions.

1. Favorable design aspects of the M-Competitions

The M-Competitions provide a model for conducting scientific research. They employ at least five key aspects: empirical testing, multiple hypotheses, large samples, independent validation, and full disclosure. While these aspects might seem obvious, studies in management science seldom include all of them.

1.1. Empirical testing

Empirical testing is necessary to test forecasting methods. Despite the resistance of time-**series** researchers (Fildes & Makridakis, 1995), interest in empirical studies has been growing among forecasters. Forecasting journals now publish many empirical comparisons. The M-Competitions have led the way in such comparisons.

1.2. Multiple hypotheses

Academic researchers rely heavily upon advocacy (Armstrong, Brodie, & Parsons, 2001b); they develop what they believe to be the beat method (hypothesis), then seek information to support it. It is uncommon in management science for a researcher to examine competing hypotheses. However, nearly 60% of empirical papers in the *Journal of Forecasting* and the *International Journal of Forecasting* tested competing hypotheses (Armstrong, 1989). The M-Competitions have been exemplary in providing open calls, thus allowing those with different approaches to participate.

1.3. Large samples

Testing should be done from large samples. However, many academic studies, including those in forecasting, do not use large samples**.** You need only to pick up the latest copies of journals to observe this. The M-Competitions were a departure from this norm. The Brat forecasting competition (Makridakis & Hibon, 1979) examined 111 series (considered large at the time) and the M-Competition (Makridakis et al., 1982) examined 1001.

1.4. Independent validation on a common data base

The methods in the competitions were tested on a common holdout database by a researcher who examined the accuracy of forecasts submitted by the competitors. This testing procedure avoided problems inherent in drawing conclusions from prior research in which databases are different.

1.5. Full disclosure

Full disclosure is important to allow others to conduct replications and extensions. Despite a consensus among researchers that replication is vital in advancing scientific knowledge, the number of published replications in the management sciences is negligible, and there are few

extensions. Furthermore, the percentage of studies in which the replications supported the original findings is low (Hubbard & Vetter, 1996).

For the most part, the M-Competitions have reported the data (forecastingprinciples.com), forecasts, and details about the methods. Replications and extensions of the M-Competitions have supported the original findings.

The M3-Competition did not require full disclosure by those using commercial packages, however, and I believe that this was a reasonable departure from the full-disclosure rule. It is to the credit of the software firms that they were willing to compete. That said, it would be difficult to determine which aspects of the commercial packages are most useful, so one cannot draw generalizations about forecasting methods from their results.

2. Suggestions for redesign of future competitions

While the M-Competition's use of empirical testing, multiple hypotheses, large samples, independent validation, and full disclosure represent a major advance in research on forecasting, improvements can be made in the approach. First, criteria besides accuracy should be examined. Second. studies should include domain knowledge. Third, studies should examine the effectiveness of specific forecasting procedures. Finally, hypotheses should specify conditions under which one might expect certain results.

2.1. Examine criteria beyond accuracy

In addition to accuracy, other criteria are important to researchers and practitioners. In Armstrong (2001), 1 describe 16 criteria, such as ability to compare different policies, reliability of confidence intervals, and ease of use, that can be used to compare forecasting methods.

2.2. Use domain knowledge

It made sense initially to simplify the problem and to assume that domain knowledge was not available. This assumption is often made in practice when forecasting thousands of items for inventory control. However, forecasters can add more value to situations in which there is domain knowledge (Armstrong & Collopy, 1998).

The original M-Competition provided some domain knowledge in the brief descriptions of the series. The M2-Competition, run in real time, gave forecasters an opportunity to draw upon domain knowledge. However**,** such knowledge was not used, perhaps due to the lack of a systematic way to use the information. Future competitions should provide information about the series so that forecasters can easily incorporate this knowledge. A structured scheme would help forecasters use domain knowledge. We proposed such a scheme, causal forces. and use it earlier on a sample of M-Competition series (Armstrong, Adya, & Collopy 2001a).

2.3. Test each procedure used in forecasting models

As Makridakis and Hibon claim in their M3-Competition paper, the goal of the competition is to "better understand the factors that affect forecast accuracy." However, the M3 design does not allow for such an assessment because each forecasting method is comprised of many procedures (e.g., adjust for seasonality, handle outliers, estimate trend) and the analyst cannot assess each procedure. To assess their impact on performance, we need to identify the various procedures used in the methods, hypothesize how they affect performance, and conduct experiments.

2.4. Include conditions in hypotheses

Researchers often fail to specify conditions in social sciences (Armstrong et al., 2001b). This applies to the M-Competitions. Prior to validation, researchers should describe the conditions under which their methods will produce better results, and describe reasons for these expectations.

To identify when certain procedures work well, the conditions for each series must be described. Armstrong et al. (2001a) list 28 descriptors, which include time interval, length of forecast horizon, causal forces, number of observations, direction of basic trend, and length of recent run. The analysts should report on performance in such a way that one can assess which procedure works best under what conditions.

With 28 descriptors, the number of possible conditions is very large. Even if guided by theory, research should employ massive databases, perhaps hundreds of thousands of series, using successive updating and multiple horizons, so that millions of forecasts can be used for development and validation. The ability to perform such studies now exists.

3. An alternative approach to competitions: Variations on a common model

Forecasting competitions can be conducted by starting with a basic model that uses the best procedures available. Guidelines could then be proposed as to what changes in procedures would be most effective under what conditions. For example, one might try alternative procedures for determining tends in situations in which uncertainty is high. The guidelines would be tested on the same data used by, the basic model. By keeping the data and all other aspects of the method constant, one could substantially reduce the need for data when testing a procedure. In addition, then would no longer be a need to coordinate the efforts of a group of forecasters. The researcher would simply compete against the model that is based on existing forecasting knowledge. I expect that this would be much less expensive than the large-scale competitions.

Rule-based forecasting (Collopy and Armstrong, 1992: Armstrong et al., 2001a) can be used to represent the best practices in extrapolation. The guidelines (rules) have been published and are posted on websites (forecastingprinciples.com). The program is available to researchers. The rule base can be modified when it is shown that a new guideline is more effective than an existing one. New models can be developed if RBF does not meet the needs of the researcher; for example, a new model would be needed to teat econometric methods. Whatever model is used, the keys are to describe or control all elements of the system except for the one that is being studied

.

References

Armstrong. J. S. (1988), "Research needs in forecasting," *International Journal of Forecasting*, 4, 449-465.

Armstrong. J. S. (2001), "Selecting methods," in Armstrong, J. S. (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell, MA: Kluwer Academic Publishers, pp. 365-386.

Armstrong. J. S. & Collopy, F. (1992), "Integrative or statistical methods and judgment for time series forecasting: Principles from empirical research," in G. Wright and P. Goodwin (Eds.), *Forecasting with Judgment*, Chichester, UK: John Wiley.

Armstrong. J. S., Adya, M. & Collopy, F. (2001a), "Rule-based forecasting: Using judgment in time-series extrapolation," in Armstrong, J. S. (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners*, Norwell, MA: Kluwer Academic Publishers, pp. 259-282.

Armstrong, J. S. Brodie, R. & Parsons, A. (2001b), "Hypotheses in marketing science: Literature review and publication audit," *Marketing Letters*, 12, 171-187.

Collopy, F. & Armstrong, J. S. (1`992), "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations," *Management Science*, 38, 1394-1414.

Fildes, R. & Madridakis, S. (1995), "The impact of empirical accuracy studies on time-series, analysis, and forecasting," *International Statistics Review*, 63, 289-308.

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

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

Makridakis, S. & Hibon M. (1979), "Accuracy of forecasting: An empirical investigation," *Journal of the Royal Statistical Society, Series A*, 142 (Part 2), 97-145.