Published in the *Journal of Business* (1978), 51 (4), 595-600.

Econometric Forecasting and the Science Court

J. Scott Armstrong
The Wharton School, University of Pennsylvania

A man convinced against his will
Is of the same opinion still.
(Paraphrased from Samuel Butler, 1663)

My paper, "Forecasting with Econometric Methods: Folklore versus Fact," argued for the method of multiple hypotheses in econometrics. However, those who favor advocacy (e.g., Mitroff 1972) assert that ample opportunity exists for minority viewpoints to be aired. Furthermore, they feel that the marketplace for ideas will work efficiently.

A recent proposal in *Science* (reviewed by Boffey 1976) called for a science court to improve the advocacy procedure (see also Callen 1976; Presidential Advisory Group 1976). Rather than wait for different viewpoints to appear in the literature, scientists would be called to act as advocates, much as in a court of law. Such a procedure should speed up the collection and examination of the evidence.

The Journal of Business has, in effect, provided a science court to examine the issues discussed in "Folklore versus Fact." Seven scientists were asked to participate in this endeavor. The *Journal of Business* assembled an impressive group of scientists, most of them well-known econometricians. These scientists also bring a variety of viewpoints with them.

What should we expect of the science court? It should at least help to ensure that all relevant evidence on the issues has been presented. A second objective, though of lesser importance, is that it make recommendations on the basis of the evidence.

The "Folklore versus Fact" paper obtained the relevant evidence by using the method of multiple hypotheses. I claimed that such an approach is efficient and relatively unbiased.

What we have then, is a modest opportunity to contrast the method of multiple hypotheses with the most desirable form of the advocacy method – that is, with the science court.

The next section of this paper examines the additional evidence produced by the seven scientists on each of the issues.[1] The issues were: (1) "Should econometricians use the method of multiple hypotheses rather than advocacy?" (2) "Do econometric methods provide the most accurate approach to short-range forecasting" (Table 2 of "Folklore versus Fact")? (3) "Are complex econometric methods more accurate than simple econometric methods" (Table 4 of "Folklore versus Fact")? For simplicity, I refer to these issues as research strategies, short-range forecasting, and complex methods, respectively.

Evidence Provided by the Seven Scientists

If the advocacy method were superior to multiple hypotheses, it should generate substantially more evidence and the evidence should be more balanced.[2] Did this happen? I examined each of the papers by the seven scientists (and

[1] The editors of the *Journal of Business* made available to me a major work by Professor Victor Zarnowitz of the University of Chicago, "An Analysis of Annual and Multiperiod Quarterly Forecasts of Aggregate Income, Output, and the Price Level,"' which, though not directed at my "Folklore versus Fact" paper, nonetheless has implications for the issues discussed therein. 1 therefore refer to that paper as well. (The Zarnowitz paper appeared in the January 1979 issue of the *Journal of Business*.)

[2] As in "Folklore versus Fact," evidence was defined as published empirical evidence.

their key references) and summarized them as follows: no = no presentation of new evidence; yes = provided new evidence. The results of this assessment are provided in Table 1.

TABLE 1 Provided Empirical Evidence on . . .

Author	Research Strategies?	Short-Range Forecasting?	Complex Methods?
Chow	No	No	No
Kosubud	No	Yes	No
McNees	No	No	Yes
Miller	No	No	No
Wecker	No	No	No
Zarnowitz	No	Yes	Yes
Zellner	No	No	No

The overriding conclusion is that the science court did not produce much new evidence. For example; no evidence was provided on the issue of the most desirable research strategy.[3]

Two pieces of evidence related to short-range forecasting:

1. Kosobud challenged our coding of his study that econometric forecasts were superior to judgmental forecasts. Inasmuch as it was his study, this was interpreted as new evidence. He does not state how he would code his study, however. I have inferred that he sees "no difference." (Incidentally, all three coders had classified Kosobud's study in an identical fashion.)

2. Zarnowitz provides evidence that econometric methods are inferior to judgmental methods. (A comparison of the median econometric model with the survey median in Table 2 of his paper indicated that the survey was more accurate on two of the three comparisons. No test of statistical significance was performed on these results.) On the other hand, the econometric forecasts were superior to the extrapolation forecasts. (This was based on four of the six comparisons in his Table 2. Again, no test was made of statistical significance.)

If this new evidence were added to Table 2 of “Folklore versus Fact,” there would be little change. Kosobud's study does not qualify, as it was an ex post forecast. Thus, the score for ex ante forecasts goes to: Seven comparisons with econometric methods superior; three comparisons with no difference; and eight comparisons with econometric methods inferior. In other words, the record for econometric methods would be about the same as in “Folklore versus Fact.”

Two pieces of evidence were added to the complexity issue: (1) McNees, referring to his 1973 paper, found a more complex model (by Fair) to be superior to a less complex model (by the St. Louis Federal Reserve Bank) for an ex post forecast; (2) Zarnowitz, in comparing three econometric models, found no differences in their accuracy. Thus there would be no relationship between the complexity of these models and their accuracy. ('This would be classified as indirect evidence.) Thus, the balance of evidence in Table 4 of “Folklore versus Fact” has not been changed.

Instead of providing new evidence, much effort was devoted to a reevaluation of the evidence in “Folklore versus Fact.” For example, a number of the scientists were critical of the survey of the experts. In terms of the three key issues, however, this point is of little consequence. The survey was merely part of the evidence suggesting that many leading econometricians believe econometric methods to be more accurate than other methods for short-range

[3] Remember, however, that “Folklore versus Fact” provided no direct evidence on this issue. The method of multiple hypotheses was proposed as a way around "selective perception." That is, multiple hypotheses should improve objectivity. A reference was also provided to a paper containing a summary of the empirical evidence on multiple hypotheses versus advocacy.

forecasting and that they, believe complexity is a virtue. Furthermore, the survey was clearly labeled as being drawn from a convenience (nonprobability) sample, not from a probability sample.

The need for a survey of experts became apparent when I first presented my findings at conferences in Europe. Econometricians at these sessions generally had one of two reactions: (1) They were not surprised at the empirical findings because they also felt the folklore to be wrong; or (2) they did not believe the empirical findings. For the latter, the results that I presented seemed to increase their belief in the folklore. According to the "I knew it would happen" research (e.g., Slovic and Fischhoff 1978), the first reaction is common. The second reaction is related to the reduction of cognitive dissonance. One implication is that those who believe strongly in a hypothesis will increase their belief in this hypothesis if they are given evidence that refutes this belief (Batson 1975). Interestingly, the better the quality of this dissonant evidence (as judged by that individual), the greater the increase of confidence in his belief.

To handle the "I knew it would happen" problem, I presented the survey at the beginning of my presentation. These results (similar to those in "Folklore versus Fact") were then used to show that the empirical evidence conflicted with the beliefs of the people at the conference.

Recommendations by the Science Court

The most valuable contribution of the science court lies in marshaling the evidence. The readers can make their own recommendations on the basis of the evidence. Still, it may be of value if the recommendations are explicitly stated by those in the science court.

Before presenting the recommendations of the science court, a summary is provided of the recommendations implied by the "Folklore versus Fact" paper:

1. The primary recommendation was that econometricians should make greater use of the method of multiple hypotheses. Such studies provide the most practical way to make comparisons among different hypotheses. (For example, Tables 2 and 4 of "Folklore versus Fact" were drawn solely from studies using multiple hypotheses.)

2. Resources in econometrics are being misallocated; there should be a shift from short-range to long-range forecasting problems.

3. Econometric forecasts can be combined with forecasts from other methods to yield improvements in accuracy.

4. Without strong evidence to suggest otherwise, econometricians should use relatively simple and inexpensive methods. (The burden of proof should be on those who propose more complex methods.)

An examination was also made of the recommendations by the members of this science court. My interpretation of these recommendations follows: *Chow*: put more emphasis on the use of econometric methods for long-range forecasting; *Kosobud:* combine short-range econometric forecasts with those from other methods; *McNees*: (1) retain econometric methods in short-range forecasting not because of accuracy, but for obtaining consistent forecasts rapidly, (2) conduct further study to identify the optimum level of complexity in the econometric methods for a given type of forecasting problem; *Miller:* (1) use tests of stability as an important way to test econometric methods, (2) continue the search for better theory and better techniques (i.e., continue with existing strategies); *Wecker*: (1) be skeptical of large-scale econometric models, (2) retain the advocacy strategy; *Zarnowitz:* do not use econometric methods for long-range forecasts; *Zellner:* use simpler econometric methods.

The point that draws the most agreement is that econometricians generally use methods that are too complex. This increases costs, inhibits communication, and does nothing to improve accuracy. In view of this, McNees's recommendation to determine the optimum level of accuracy seems to be of particular value.

Some research has been done to determine the optimum level of complexity in econometric methods. A review of this evidence is provided in Armstrong (1978). Einhorn and Hogarth (1975) was one of the more useful studies discussed in this review. Recent studies by Newman (1977) and Wainer (1976) should be added to the studies cited in my review. This evidence suggests that it is important to select a small number of causal variables and to specify the direction of the relationships in a model. Surprisingly, complexity beyond this point seldom leads to improved accuracy in practical problems. Under certain conditions—namely, (1) large sample sizes, (2) accurate measurement of the dependent variable, (3) negative intercorrelations among causal variables (ones that cannot be eliminated by rescaling), and (4) strong causal relationships (such that the causal variable will cause large changes over the forecast horizon)-it helps to obtain regression estimates of the causal relationships. Otherwise, unit weights are sufficient. No real-world conditions have yet been identified where complexity beyond ordinary least-squares methods would yield improvements in ex ante forecasts.

The fourth condition, large changes due to causal factors, is of particular importance. This helps to explain why econometric methods have done well for long-range forecasting but not for short-range forecasting.

Conclusions

The science court provides safeguards. Because most of the seven scientists took exception to the conclusions in "Folklore versus Fact," one gains confidence that the evidence against these conclusions has been fairly presented. However, little new evidence was generated.

The science court produced a number of recommendations – some agreeing with those in "Folklore versus Fact," and others disagreeing. The point of most agreement was that econometricians should use simpler methods.

In my opinion, the science court provides a useful procedure to increase the value of advocacy. Note, however, that the method of multiple hypotheses, as employed in "Folklore versus Fact," led to an efficient and apparently unbiased collection of the evidence (additional evidence by the science court tended to support the conclusions in "Folklore versus Fact"). And the collection of evidence is a key function of the scientists. Once this is done, the scientist can promote his findings but this can also be done by the readers, who are the ultimate judges in our science court.

References

Armstrong, J. Scott (1978), *Long-Range Forecasting: From Crystal Ball to Computer*. New York: Wiley-Interscience. (available in full text at forecastingprinciples.com)

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

Boffey, Philip M. (1976), "Experiment planned to test feasibility of a science court," *Science*, 193, 129.

Callen, Earl (1976), Letter, *Science*, 193, 950-951.

Einhorn, Hillel J. and Hogarth, R. M. (1975), "Unit weighting schemes for decision making," *Organizational Behavior and Human Performance*, 13, 171-192.

Mitroff, Ian (1972), "The myth of objectivity, or why science needs a new psychology of science," *Management Science*, 18, B613-B618.

Newman, J. Robert (1977), "Differential weighting in multiattribute utility measurement: When it should not and when it does make a difference," *Organizational Behavior and Human Performance*, 20, 312-325.

Presidential Advisory Group (1976), "Report by the Task Force of the Presidential Advisory Group on Anticipated Advances in Science and Technology," *Science*, 193, 653-656.

Slovic, Paul, and Fischhoff, Baruch (1977), “On the psychology of experimental surprises,” *Journal of Experimental Psychology: Human Perception and Performance*, 3, 544-551.

Wainer, Howard (1976), “Estimating coefficients in linear models: It don't make no never-mind,” *Psychological Bulletin*, 83, 213-217.