PROCEEDING PAPERS

WINTER MEETING OF THE AMERICAN AGRICULTURAL ECONOMICS ASSOCIATION
WITH ALLIED SOCIAL SCIENCE ASSOCIATIONS

Toronto, December 28–30, 1972

THE VALIDITY AND VERIFICATION OF COMPLEX SYSTEMS MODELS

CHAIRMAN: SAUL H. HYMANS, UNIVERSITY OF MICHIGAN

Is Verification Possible? The Evaluation of Large Econometric Models*

HAROLD T. SHAPIRO

CONSTRUCTION of large econometric models began in earnest during the years of World War II, although some earlier pioneering examples are available. The past two decades, however, mark the period of relatively large and sustained expenditures of energy and resources on the construction of these models. Currently, large econometric models are daily becoming more numerous, more "sophisticated," larger, and certainly more popular. This rising tide of popularity has been temporarily checked from time to time due to the occasional gross forecasting errors generated by these models, but on the whole these "credibility gaps" have been short-lived, and the increasing popularity of these devices has flowed on much like GNP and "Ole Man River." Use of large-scale econometric models for forecasting is now widespread in both government and industry. In addition, these models are used by government policy makers for the calculation of policy multipliers, the evaluation of old policies, and for the ranking of prospective new ones. Despite the central importance some of these models have assumed in decision making, the process of systematic model evaluation seems, with some noticeable exceptions, to have received relatively little attention. At the very least the problem of model evaluation has lagged behind the frenetic activity in the area of model construction. Important early work in model evaluation was done by Christ [6], Adelman [1], and Theil [28], among others, and more recently the studies of Zarnowitz, Boschan and Moore [36], Evans, Haitovsky, and Treyz [11], Haitovsky and Wallace [16], as well as the current NSF–NBER Seminar on the Comparison and Evaluation of U. S. Econometric Models [9], [12], and the recent (1969) NBER Conference on Research in Income and Wealth [17], have added considerably to our understanding of this area. Nevertheless it remains true that we still lack a clear and accepted analytical basis for the selection of proper criteria for model evaluation.[1] Indeed, except for the simplest cases (nested hypothesis) the problem of evaluation and verification of single equation models is far from solved. In this latter area, although there are many more techniques available to aid in the process of model evaluation, there seems to be, among economists, no general agreement on the meaning and purpose of model verification.

The purpose of this paper is to help clarify the notion of model verification, particularly in the empirical sciences and suggest how these considerations might "shape" our approach to the problem of the verification of large econometric models. I conclude that it is not possible to consider establishing either the truth or falsity of any theory about the structure of our economic environment, or the models that represent it, and we should instead, turn our attention to the purpose of economic theories

* Portions of this paper were originally prepared for the Seminar on Criteria for Evaluation of Econometric Models (S. H. Hymans and H. T. Shapiro, co-chairmen) of the NBER-NSF Conference on Econometrics and Mathematical Economics. I am grateful to members of this seminar for many helpful ideas and criticisms.

HAROLD T. SHAPIRO *is a professor of economics and Director of the Research Seminar in Quantitative Economics at the University of Michigan.*

[1] This state of affairs was, in part, the motivation for the NBER-NSF seminar on Criteria for Evaluation of Econometric Models. For this seminar's statement of the "state of the art," see Dhrymes, *et al.* [9].

and from these objectives develop procedures for the evaluation of econometric models. The section below considers the problem of verification in the empirical sciences; the final section confronts the issue of model verification with respect to large econometric models. Hopefully the discussion will help improve our understanding of the implications of the battery of statistical techniques we have been using.

Verification in the Empirical Sciences[2]

A scientist, whether he is a theorist or an experimenter ("verifier" or tester), is concerned with putting forward statements (hypotheses, propositions) of various types and then testing them step by step. This paper, however, is not concerned with the analysis of the act of conceiving a new idea, or just how it happens that a new scientific theory occurs to someone, but simply with one particular aspect of the overall procedure of scientific inquiry—the process of systematically testing new ideas. All new ideas, of course, must be subjected to some set of systematic tests if they are to be seriously entertained. The nature and the implication of these tests, however, differ markedly in the empirical sciences (e.g. economics) and the nonempirical sciences (e.g., formal mathematics and logic).

Testing in the empirical sciences is concerned with confronting what we may call empirical theories, or theories that can be characterized as a set of propositions which are conceivably capable of being tested by observation and/or experience. The propositions of an empirical theory, therefore, are composed of *synthetic* statements or statements whose validity depends on the facts of experience. The propositions of formal mathematics or logic on the other hand are composed of *analytic* statements, or statements that convey no information about the world of experience. These propositions are necessarily true (if free from logical contradictions) since they have no factual content. The validity of nonempirical theories therefore depends solely on definitions of the symbols they contain and the rules of deductive logic. For example, consider the following statement: "either some firms are monopolies or none are."[3] Such a statement conveys no information at all about firms, and thus no conceivable experience can refute it. This should not be taken to mean that all *analytic* statements are useless. On the contrary, they very often reveal usages and relationships which we might otherwise not be conscious of. As a simple example, consider the following proposition: "If all households are net savers and all net savers accumulate capital, then all households accumulate capital." Such a statement, while not logically refutable by any conceivable set of facts, does point to the derivable relation that all households accumulate capital. From the point of view of testing new theories or ideas, the key distinction between empirical and nonempirical theories is that the former may be free from logical contradiction and still be false. That is, empirical theories may fail, not because they are formally defective, but because they fail other types of criteria based on observation and/or experience.[4]

Even in the empirical sciences, however, it is necessary to draw a distinction between practical testing and testing in principle. There are many empirical propositions which we understand and in many cases believe, even though they have not in fact received any systematic testing. Many of these could be tested if we took the trouble, but others we could not test appropriately, even if we chose, as we lack the practical means of placing ourselves in the situation where relevant observations can be made. Thus, although it is not practically possible to test such propositions at the moment, we can specify what set of observations would be necessary to carry out our test, once we are in a position to make such observations. The proposition that, "the inner core of Jupiter is white hot," is an example of a statement that is testable only in principle since an actual test must await our ability to bore into the interior of that planet. Nevertheless this theory on the temperature of Jupiter's core qualifies as an empirical theory as its propositions are, in principle, testable by observation and experience.

In considering the appropriate set of procedures to test the propositions of an empirical theory, the first step is to ensure that the theory (set of propositions) being advanced does in fact

[2] For a more systematic and learned discussion of this issue see Ayer [2] and Popper [27].

[3] This statement has the logical form "either X is true or X is not true."

[4] Metaphysical theories, of course, are not capable of nor in need of any type of testing. Consider the following metaphysical statement: "the Absolute enters into, but is itself incapable of evolution and progress" (F. H. Bradley, *Appearance and Reality*). One cannot conceive of either an observation which would enable one to determine whether or not the Absolute did or did not enter evolution and progress, or a logical contradiction of that statement. Such a statement is not even in principle verifiable.

fulfill the requirement of such a theoretical system. It should contain the following characteristics:

(1) It should be composed of synthetic, rather than metaphysical or analytical statements. That is, the theory should represent both a noncontradictory and possible world (i.e., a world of possible experience) where the validity of each of the theories or propositions may be tested by the facts of experience. That is, it must be possible for the ideas advanced to be refuted, in some sense, by experience.

(2) The theoretical system must be distinguishable from other systems that represent the world of experience in the sense that there will be a recognizable scientific advance should the propositions "pass" the various tests.

The initial stage of testing a new theory consists of examining its various propositions along with other relevant statements to determine what logical relations exist between them such as equivalence, compatibility, and incompatibility. The logical form of the theory is then examined to determine whether it is an empirical, nonempirical, or metaphysical theory. Should all be in order at this stage, we can then begin actual testing by way of empirical applications of the conclusions (statements) which form part of or have been derived from the new theory. This next stage involves facing the problem of model verification, or the problem of deciding whether or not the model *truly* represents some specific aspect of the world of experience; when we conceive of the task of model verification as the problem of establishing the truth of a particular set of synthetic statements (empirical theory), we must next specify what set of observations will accomplish this task. However, since our world of observation and/or experience is necessarily finite, we are faced with the well-known "problem of induction," or the problem of deciding when, in the empirical sciences, singular statements (summaries of our observations) can lead to universal ones (those of our theories). It seems clear that one is never justified in inferring universal statements from singular ones regardless of how numerous are the latter. No matter how many times we observe that increases in the supply of money are inflationary, the evidence will never justify, *in a formal sense*, the conclusion that all increases in the money supply are inflationary. Similarly, no matter how many white swans we see, it does not justify the statement that all swans are white. However strong the evidence in favor of a particular set of propositions, there is never a point at which it is impossible for further experience to go against it. None of the empirical propositions of a theory are absolutely certain in the sense that their truth either has been or may be absolutely confirmed. Science, simply does not have the power to decide the truth of its statements. Moreover, empirical science cannot, on the basis of observation, establish the falsity of any empirical theory.[5] The "facts" of experience can never, on logical grounds, compel us to abandon a favorite hypothesis. Suppose we have devised an experiment to test the validity of a certain theory. All theoretical propositions state that under certain conditions a certain type of event (or observation) will occur. Thus when we make the observation predicted by the theory, it is not only the particular law that is substantiated but the existence of the requisite conditions. If we fail to make the predicted observations, we may say that conditions were not what they seemed to be and may construct a theory to explain how we were mistaken about them. Any particular instance in which a favorite hypothesis appears to be refuted can always be explained away (particularly, as we shall see below, when dealing with nonexperimental models such as econometric models), although the scientist must be careful to retain the possibility that the hypothesis may ultimately be abandoned under the "force of circumstances," or the hypothesis can no longer be considered a genuine part of an empirical theory. Any empirical proposition we are resolved to maintain in the face of any experience is not a *synthetic*, but an *analytic* proposition.

If empirical theories cannot be verified, we may also wonder how we can speak of objective scientific statements in the empirical sciences—i.e., statements that are independent of a par-

[5] There is a class of rather uninteresting empirical propositions which can be verified conclusively by observations. These are propositions which simply convey the content of a single experience, e.g., "The sun rose yesterday." However, these types of propositions hardly constitute a theory. Similarly there is a class of empirical theories whose falsity can, in principle, be established. These are collections of propositions put forward in the following way: "Under every conceivable set of conditions the following exact (nonstochastic) relationship will hold. . . ." In this latter case a single counter-example establishes the falsity of the theory. Neither of the two classes of theories, however, has any importance for the advancement of empirical science.

ticular investigator's whim. The objectivity of empirical theories consists only in the requirement that they be "intersubjectively testable." Thus in the empirical sciences, scientific objectivity is simply interpreted as a methodological rule. Before a set of statements and tests can be eligible for introduction into a science, it must be possible to issue a precise set of instructions that will enable other researchers to reproduce the results exactly.

If science cannot decide on the truth or falsity of its theories, how do we determine which hypotheses shall be "retained" (used to guide action) and which shall be abandoned? The answer is readily available if we consider the reason for construction of theories and/or hypotheses in the first place. People may construct theories for the "fun of it," but scientists construct theories in order to help us better anticipate or forecast some aspect of our world of experience. The principle objective of new empirical theories is to enable us to make more accurate predictions. To put the matter another way, empirical theories can be thought of as rules which govern our expectation of future events. What I would suggest, therefore, is that the validity of an empirical hypothesis can be "tested" (not verified) by seeing whether it actually fulfills the function for which it was designed. An experiment designed to "test" a particular proposition can be considered a "success" and the hypothesis corroborated if the results increase our willingness to use the hypothesis as a guide to future events or as a guide to action. The scientist's desire to have an efficient set of rules for prediction induces him eventually to take notice of unfavorable observations and abandon cherished hypotheses. Thus logically constructed models and the theories they represent cannot be verified in the empirical science, nor can their falsity be established. We can, however, think of a theory as temporarily corroborated if the tests to which it has been subjected increase the degree of confidence with which we await the fulfillment of "the prediction."[6] Therefore we should not speak of verification of econometric models but of continuous evaluation or corroboration of models whereby such models are subject to a series of evaluative tests specified in the light of the particular uses to which it is desired to put the model. At best we ought to think of a model as being *tentatively* certified as a reasonable tool until it generates an error serious enough to shake our confidence or until it is replaced by a better unverified ("untrue") model. In this context evaluation becomes a problem–dependent or decision–dependent process, differing from case to case as the proposed use of the model under consideration changes. Thus a particular model may be "validated" for one purpose and not for another. In each case the process of evaluation is designed to answer the question: Is this model fulfilling the stated purpose? We can then speak of the evaluation of these models as the process of attempting to validate them for a series of purposes. Thus the motivation of model–builders or users becomes directly relevant to the evaluation of the models themselves. The "success" of a model can be measured by the extent to which it enables its user to decrease the frequency and consequences of wrong decisions. As Zarnowitz [34] has pointed out, the full application of even this more limited goal still poses very high informational requirements: the errors must be identifiable, the preferences of the decision maker and the constraints under which he operates must be available, and the cost of providing the model must be ascertained.

The Evaluation of Large Econometric Models

Construction and evaluation of large econometric models can be thought of as composed of two *interrelated* stages—formulation and estimation of model components and evaluation of the model as a whole. The first stage consists of a number of distinct activities. To begin with there is the formulation of a set of working hypotheses about the behavior of the system being modeled. At this point we use all available information—observations, general knowledge, relevant theory, and intuition. This stage also includes the specification of variables and functional relationships, sample selection, and the selection of appropriate estimation procedures.[7] In principle all specification should be done at this stage, but *a priori* information may not be

[6] Needless to say, the increased degree of confidence generated by any "successful" test of a theory will depend on both the severity of the test (and previous ones) and the "degree of testability" of the theory (probability of being refuted). Nothing is simpler than to construct a theoretical system which is consistent with a certain set of facts, but this will not, by itself, fulfill the function of theories suggested above.

[7] Howrey, *et al.* [19] pointed out that the method of estimation itself may also be partially a function of the use to which the model is to be put. The evaluation of any model should, of course, include an evaluation of the estimating procedures used. We do not comment on this aspect of the evaluation process here. For an interesting discussion of this issue see [19].

rich enough to allow us this luxury. Some of the processes of model specification may have to be left until the second stage of the procedure (overall model evaluation). As with specification, the final determination of sample–size and estimation procedures may come in one of the later stages of the model building process as these latter procedures interact with the specification process. Finally, this stage also includes attempts to validate the individual hypotheses on which the model will be based subject to the limitations of existing statistical tests, and final decisions regarding the method of estimation. As noted above, there are many unresolved problems regarding the appropriate procedures (including evaluation) in the initial stage, but this paper does not address these problems. Rather, we are more directly concerned with the second stage which considers procedures for the evaluation of the model as a whole.

Given that model evaluation is a problem–dependent process, we must begin the process of evaluation by considering the uses to which the model will be put. For example, one may accept the Box–Jenkins purpose of model building: "Parametric modeling attempts to discover the structure of a time series so that the residuals, after fitting the model, are purely random or white noise" [3]. Their suggested three–stage iterative procedure of *specification*, *estimation*, and *residual analysis* is exhaustive. For most economists, however, this approach would seem too narrow.[8] For most problems which economists concern themselves with and for which they will seek the help of large econometric models, it seems hard to escape the conclusion that the purpose of these models, one way or another, is to predict some aspect of reality. Thus we are bound, at least initially, to be concerned with either the model's retrospective predictions (historical validation) or prospective predictions (forecasting). Even where the model is being used for policy analysis with no historical counterpart to the simulated series (except for the "control" solution), our confidence in the result (our willingness to use the results for decision–making) will almost certainly depend on the ability of the model to predict both within the sample and in the post–sample period. Post–sample simulations are especially important for the purpose of hypothesis testing, as opposed to the procedures of hypothesis searching which often characterize the "data mining" of the sample period. Our tests or evaluation procedures should initially center on the ability of the data generated by "historical" simulation experiments to conform to the actual data. These simulations might be either deterministic or stochastic and either static (one period) or dynamic (multiperiod) in nature. A minimal requirement would be the comparison of the simulated data generated by a deterministic single–period simulation with the data from the actual historical record (both within and outside the sample period).

However, even if a model "passed" a somewhat more demanding test of its ability to "track" the historical record (e.g., a deterministic multiperiod historical simulation), economists normally also want to investigate whether or not the model responded to various types of stimuli in the fashion anticipated or suggested by economic theory or independent empirical observation. Quite aside from the individual hypotheses underlying particular equations in the system, economists have certain (not entirely independent) *"reduced form" hypotheses* which they would demand "acceptable" models to conform to. As a profession we seem to have developed some more or less vague ideas about the magnitudes of various impact, dynamic, and steady–state multipliers as well as some prior notions about other dynamic characteristics that the model "should" exhibit. Despite Haavelmo's early warning [14], however, we have, at least until the recent work of Howrey [20], failed to realize just how difficult such tests are to design and carry out. Almost all model evaluation procedures to date have employed nonstochastic simulation (with respect to both the equation error term and the sampling distribution of the estimated parameters) to generate the experimental data—a procedure which is inadequate in testing dynamic theories. This set of issues was partly confronted again at a recent NBER conference concerned with whether or not an existing set of models reproduced adequately the cyclical swings observed in our economic system.[9] It is difficult to catalogue what seems to be a minimal set of demands of this sort as needs and requirements

[8] This paper does not consider directly the important "optimal predictor" theories of Box and Jenkins [3], Wiener–Kalmogorov [31], and Kalman and Bucy [23], or the various techniques of exponential smoothing [4]. Although these techniques have important uses, it does not seem to this author that at the current time they can have an important role in the evaluation of economic hypotheses.

[9] Conference on Research in Income and Wealth, Harvard University, November 14–15, 1969. For a summary introduction to these issues as they arose at this conference, see [17].

vary according to the preferences and prejudices of the researcher and the actual needs of the user. In any case, constraints imposed by these demands are, given the current state of knowledge, not overly stringent. Even if we consider the case of the government expenditure multiplier, where a relatively large amount of evidence has accumulated, "acceptable" estimates of its magnitude (both impact and steady-state) vary widely among different "accepted" models of the U. S. economy (see Fromm and Klein [12]). I will devote most of my attention, therefore, to the use of predictive tests in the evaluation of these models.

We should also briefly consider, however, whether in all types of experiments the simulated data should be generated by stochastic or nonstochastic simulation procedures. Certainly stochastic simulation, if we have the necessary extra information (in practice we often ignore the problem of obtaining good estimates of the variance–covariance matrix of the disturbance process), will yield a more precise characterization of the model being used and thus increase the quality of the evaluation procedure. Further, if the model is nonlinear (most are these days), then the reduced form of the model is *not* the same as the nonstochastic solution [18]. Thus application of nonstochastic simulation procedures yields results that should not be expected to be consistent with the properties of the actual reduced form of the model. The question of whether this difference is large or not remains to be tested. Preliminary experiments with the Wharton model suggest the difference is not great, but a more recent study by Haitovsky and Wallace [14] suggests the contrary. When feasible, it seems advisable to use stochastic simulation to generate the experimental data.

Evaluation of the predictive ability of a model is essentially a goodness of fit problem. Although there are a number of well developed statistical techniques ostensibly available for this purpose (many initially developed in the experimental design literature), econometric model builders have mainly restricted themselves to simple graphical techniques (the fit "looks good") or simple summary measures (root mean square error, Theil's U Statistic)[10] of the performance of certain key variables (i.e., GNP). In a more recent paper, Haitovsky and Treyz [15] proposed a very interesting descriptive decomposition of the forecast error for an endogenous variable in a large econometric model. The decomposition identifies error components involving: (a) the structural equation explaining the variable in question, (b) the rest of the estimated structural system, (c) incorrect values of lagged endogenous variables (in the case of dynamic simulations), (d) incorrect guesses about exogenous variables (in the case of an *ex ante* forecast), and (e) failure to make serial correlation "adjustments" for observed errors. In addition, a recent study by Muench, Rolnich, Wallace, and Weiler [25] provides a framework for the study of the *ex post* forecast distributions of the reduced form of large non-linear econometric models and provides a valuable new addition to our evaluative procedures. Some attention has also been given to the development of a statistic analogous to the single-equation R^2, to be used to test the hypothesis that $\beta = 0$, where β is the coefficient vector of the system of equations under consideration. An interesting and complete discussion of this issue can be found in Dhrymes [10, ch. 5]; he defines such a statistic but finds that it is dependent on the unknown covariance parameters of the joint distribution of the error terms of the system. While Dhrymes also derives an alternate test procedure regarding the goodness of fit of the reduced form, this procedure involves the restriction that the number of variables in the model (endogenous and exogenous) be less than the total number of observations—a restriction not generally fulfilled by large econometric models. The *trace correlation* statistic suggested by Hooper (based on estimates of canonical correlations) is closely related to the statistic suggested by Dhrymes, but its distribution seems quite untractable although Hooper has given an approximate expression for the asymptotic variance of the statistic. In a paper presented at the European meetings of the Econometric Society in Budapest, 1972, Carter and Nagar [5] suggested a new measure which makes use of the covariance structure of the reduced form disturbances and whose asymptotic distribution is known. This latter measure is an interesting and useful addition, but many problems still remain.

The principle technique suggested in the experimental design literature for testing goodness of fit of actual data to simulated data is the Analysis of Variance. There are, however, a

[10] Howrey, *et al.* [19] recently suggested some difficulty with the root mean square error statistic (where small sample properties are known), particularly when used to compare structural versus autoregressive models, or sample versus post sample performance of a given model. See also [9, section III].

number of difficulties in applying these techniques to data generated by large econometric models.

(a) The associated *F* and *Chi*-square tests depend on normality, constant variance, and *statistical independence*, conditions not generally satisfied by the data series generated by econometric models.

(b) Full factorial design for large nonlinear models (including all levels of the exogenous variables and the parameters describing distribution of the error term) quickly produces an unmanageably large number of cells if more than a few factors are investigated.

(c) It may be difficult to establish unbiased categories for classification of data. Occasionally a "natural" set of categories suggests itself. (See Zarnowitz's analysis of turning point errors [33].)

(d) In most interesting applications with large econometric models, we have what is known as a "multiple response problem." That is, we are interested in more than one characterization of the outcome of the experiment. If one wishes to use the analysis of variance, there are only two possibilities: treat the outcome as one of many experiments each with a single response, or combine all the responses (endogenous variables of interest) into a single response. This latter procedure, of course, involves the explicit formulation of the utility function of the user—a difficult but perhaps healthy situation. (For an interesting attempt to solve the multiple response problem see Fromm and Taubman [13] and Theil [28, 30].)

Other techniques common in experimental design literature are regression analysis and spectral analysis. With simple regression analysis we simply regress actual values on the predicted values of a series and test whether the resulting equations have "zero" intercepts and slopes not significantly different from unity (see Cohen and Cyert [7] and Hymans [21]). This general technique has also been extensively used by Theil [28], but as usual he has extended it and forced it to yield additional information. By regressing predicted values on actual values and actual values lagged a period, Theil is also able to investigate if predicted changes tend to be biased toward recent actual changes or not. Theil's inequality coefficient and its decomposition into elements of bias, variance, and covariance is very closely related to this type of analysis (although it refers to a regression of actual *changes* on predicted *changes*) and offers a great deal more information including some information on the tendency of the model to make turning point errors. Mincer and Zarnowitz [24] have provided some further development of Theil's procedure and have also suggested an additional measure of forecast error—the relative mean squared error. The latter is particularly interesting by virtue of its attempt to compare the costs and benefits of forecasts derived from alternative models of the economic process.

Spectral (cross–spectral) analysis is a statistical technique that can be used to obtain a frequency decomposition of the variance (covariance) of a univariate (bivariate) stochastic process. There are several ways in which spectrum analytic techniques might be used in the evaluation of econometric models. Naylor, *et al.* [26] suggest that the spectra estimated from simulated data be compared with the spectra estimated directly from actual data. Howrey [20] has pointed out that for linear models the spectrum implied by a model can be derived directly from the model and the stochastic simulation of the model is therefore not needed to make this comparison. Another application of spectral techniques is to test estimates of the structural or reduced–form disturbances for serial correlation, an important step in the Box–Jenkins modeling procedure.

Cross–spectral analysis can be used to investigate the relationship between predicted and actual values. That is, Theil procedures can be extended to the frequency domain using cross–spectrum analysis. This permits statistical testing of more general hypotheses about the relationship of actual and predicted values.

An important advantage of spectral analysis is that it is a nonparametric approach to data analysis. Thus it is a particularly useful device in situations in which little prior knowledge is available about the relationships under investigation. In addition, spectral methods do not depend on the statistical independence of the generated data points; they only require that the process generating the data be stationary to the second order. In order to discriminate among a number of similar hypotheses, a large number of observations may be required. Moreover, significance tests that are available depend on the assumption of normality of the underlying process or on a sample size that is large enough so that a form of the central limit theorem can be envoked. What little empirical

experience has been accumulated in connection with the use of spectral analysis to investigate econometric models suggests that the technique can be used quite effectively to investigate certain dynamic properties of econometric models.

By way of summarizing this necessarily broad discussion, I would like to present, in outline form, the range of descriptive measures which have been found to be useful in assessing the forecasting performance and other relevant characteristics of large scale econometric models. While some of these measures can be subjected to classical statistical tests, most are—at this stage of our knowledge—merely descriptive and geared to specialized model uses. Many of these procedures can be traced to the writings of Zarnowitz and his co-workers [33, 34, 35, 36], Evans, Haitovsky and Treyz [11], Box and Jenkins [3], and Theil [28].

An Outline of Nonparametric Measures[11]

A. Single-variable measures

(1) Mean forecast error (changes and levels)
(2) Mean absolute forecast error (changes and levels)
(3) Mean squared error (changes and levels)
(4) Any of the above relative to:
  (a) the level of variability of the variable being predicted
  (b) a measure of "acceptable" forecast error for alternative forecasting needs and horizons

B. Tracking measures

(1) Number of turning points missed
(2) Number of turning points falsely predicted
(3) Number of under– or overpredictions
(4) Rank correlation of predicted and actual changes (within a subset of 'important" actual movements)
(5) Various tests of randomness
  (a) of directional predictions
  (b) of predicted turning points

C. Error decompositions

(1) Comparison with various "naive" forecasts[12]
(2) Comparison with "judgmental," "consensus," or other noneconometric forecasts
(3) Comparison with other econometric forecasts

D. Cyclical and dynamic properties

(1) Impact and dynamic multipliers
(2) Frequency response characteristics.

The measures outlined have been found to be suitable for a wide variety of purposes, and surely a user's confidence in any particular model would grow in proportion to the number of positive results yielded by such of these measures as seem relevant to the use in question. Several recent studies, [16, 22], and especially the Cooper–Jorgenson study [8], have made a valuable contribution by standardizing both the period of fit and the technique of estimation across alternative models prior to conducting intermodel comparisons.[13] Further, a recent paper by Fromm and Klein [12] summarizes the work of the ongoing NBER-NSF seminar in The Comparison of Econometric Models and provides both new information and new ideas in this area.

Models will be used for decision making, and therefore their evaluation ought to be tied to optimization of these decisions. The question we have to ask ourselves, then, is what series of tests and/or procedures will be sufficient to achieve a particular level of confidence in the use of a model for a certain specified purpose?

[11] This outline is taken from [9].

[12] The procedures of Box and Jenkins [3] may be particularly powerful in helping to identify the autoregressive procedures which would best serve as "naive" alternatives to a structural model.

[13] See [19] for arguments regarding the controls needed in such standardization attempts.

References

[1] ADELMAN, I., AND F. ADELMAN, "The Dynamic Properties of the Klein–Goldberger Model," *Econometrica* 27 (1959).

[2] AYER, ALFRED JULES, *Language, Truth and Logic*, New York, Dover Publications, Inc., 1952.

[3] BOX, G., AND G. JENKINS, *Time Series Analysis; Forecasting and Control*, San Francisco, Holden-Day, 1970.

[4] BROWN, R. G., *Smoothing, Forecasting and Prediction of Discrete Time Series*, Englewood Cliffs, New Jersey, Prentice Hall, 1963.

[5] CARTER, R. A. L., AND A. L. NAGAR, *A Measure of Correlation for Simultaneous Equation Systems*, paper presented at the European meetings of the Econometric Society, Budapest, 1972, mimeo.

[6] CHRIST, CARL F., "A Test of an Econometric Model for the U. S., 1921–1947," in Universities–National Bureau Committee for Economic Research *Conference*

on Business Cycles, New York, National Bureau of Economic Research, 1951, pp. 35-107.

[7] COHEN, KALMAN J., AND R. M. CYERT, "Computer Models in Dynamic Economics," *Quart J. Econ.* 75:112-127, Feb. 1961.

[8] COOPER, R. L., AND D. W. JORGENSON, "The Predictive Performance of Quarterly Econometric Models of the United States," in *Econometric Models of Cyclical Behavior*, ed. B. Hickman, Conference on Research in Income and Wealth, Vol. 36, National Bureau on Economic Research, 1972.

[9] DHRYMES, PHOEBUS, J., *et al.*, "Criteria for Evaluation of Econometric Models," *Annals of Economic and Social Measurement*, Vol. 1, No. 3, July 1972, p. 291.

[10] DHRYMES, PHOEBUS J., *Econometrics*, New York, Harper and Row, 1970.

[11] EVANS, MICHAEL K., Y. HAITOVSKY, AND G. TREYZ, "An Analysis of the Forecasting Properties of U. S. Econometric Models," in *Econometric Models of Cyclical Behavior*, ed. B. Hickman, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[12] FROMM, GARY, AND LAWRENCE R. KLEIN, "A Comparison of Eleven Econometric Models of the United States," paper presented at the AEA meetings, Toronto, Dec. 1972, mimeo.

[13] FROMM, GARY, AND P. TAUBMAN, *Policy Simulations with an Econometric Model*, Washington, D. C., The Brookings Institution, 1968.

[14] HAAVELMO, T., "The Inadequacy of Testing Dynamic Theory by Comparing Theoretical Solutions and Observed Cycles," *Econometrica*, Oct. 1940.

[15] HAITOVSKY, YOEL, AND G. TREYZ, "The Decomposition of Econometric Forecast Error," mimeo.

[16] HAITOVSKY, YOEL, AND N. WALLACE, "A Study of Discretionary and Nondiscretionary Fiscal and Monetary Policies in the Context of Stochastic Macro-Econometric Models," in ed. V. Zarnowitz, *The Business Cycle Today*, National Bureau of Economic Research, 1972.

[17] HICKMAN, BERT G., "Introduction and Summary," in *Econometric Models of Cyclical Behavior*, ed. B. Hickman, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[18] HOWREY, E. PHILIP, AND H. H. KALEJIAN, "Computer Simulation Versus Analytical Solutions," in ed. T. H. Naylor, *The Design of Computer Simulation Experiments*, Durham, N. C., Duke University Press, 1969.

[19] HOWREY, E. PHILIP, L. R. KLEIN, AND M. D. MCCARTHY, *Notes on Testing the Predictive Performance of Econometric Models*. Dept. of Econ. Discussion Paper 173, Wharton School, University of Pennsylvania, 1970.

[20] HOWREY, E. PHILIP, "Dynamic Properties of a Condensed Version of the Wharton Model," in *Econometric Models of Cyclical Behavior*, ed. B. Hickman, Conference on Research in Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.

[21] HYMANS, SAUL H., "Prices and Price Behavior in Three U. S. Econometric Models," paper prepared for the Conference on the Econometrics of Price Determination, Washington, D. C., October 30-31, 1970.

[22] JORGENSON, D. W., J. HUNTER, AND M. NADIRI, "The Predictive Performance of Econometric Models of Quarterly Investment Behavior," *Econometrica* 38: 213-224, March 1970.

[23] KALMAN, R. E., AND R. S. BUCY, "New Results in Linear Filtering and Prediction Theory," *J. Basic Engineering*, Series D 83, 1961.

[24] MINCER, JACOB, AND V. ZARNOWITZ, "The Evaluation of Economic Forecasts," in ed. J. Mincer, *Economic Forecasts and Expectations: Analyses of Forecasting Behavior and Performance*, National Bureau of Economic Research, 1969.

[25] MUENCH, T., O. ROLNICH, A. WALLACE, AND N. WEILER, "Tests for Structural Change and Prediction Intervals for the Reduced Forms of Two Structural Models of the U.S.: the FRB-MIT Model and Michigan Quarterly Models," Research Dept. Staff Rep. (WP-19), Federal Reserve Bank of Minneapolis.

[26] NAYLOR, THOMAS H., K. WERTZ, AND T. H. WONNACOTT, "Spectral Analysis of Data Generated by Simulation Experiments with Econometric Models," *Econometrica* 37:333-352, April 1969.

[27] POPPER, KARL R., *The Logic of Scientific Discovery*, New York, Basic Books, 1959.

[28] THEIL, HENRI, *Economic Forecasts and Policy*, Amsterdam North-Holland Publishing Co., 1961.

[29] ———, *Principles of Econometrics*, New York, John Wiley and Sons, 1971.

[30] ———, *Applied Economic Forecasting*, Rand-McNally, 1966.

[31] WEINER, N., *Extrapolation, Interpretation and Smoothing of Stationary Time Processes*, Cambridge, Mass., The M.I.T. Press, 1949.

[32] ZARNOWITZ, VICTOR, "Forecasting Economic Conditions: The Record and the Prospect," in ed. V. Zarnowitz, *The Business Cycle Today*, National Bureau of Economic Research, 1972.

[33] ———, *An Appraisal of Short-Term Economic Forecasts*, National Bureau of Economic Research, 1967.

[34] ———, "New Plans and Results of Research in Economic Forecasting," *Fifty-first Annual Report*, National Bureau of Economic Research, 1971, pp. 53-70.

[35] ———, "Prediction and Forecasting: Economic," *International Encyclopedia of the Social Sciences*, The Macmillan Co. and The Free Press, 1968.

[36] ———, C. BOSCHAN, AND G. H. MOORE, with the assistance of JOSEPHINE SU, "Business Cycle Analysis of Econometric Model Simulations," in *Econometric Models of Cyclical Behavior*, ed. B. Hickman, Conference on Research on Income and Wealth, Vol. 36, National Bureau of Economic Research, 1972.