CHAPTER FIVE

VALIDATING INTERNATIONAL RELATIONS FORECASTS TO DEVELOP THEORY

Charles F. Hermann, Warren R. Phillips, & Stuart J. Thorson

I. INTRODUCTION

Periodically, policy makers, the public, and other scholars urge international relations researchers to cast the results of their studies in terms of forecasts or expectations about the future. The reasons seem clear enough. We all have an interest in anticipating aspects of future global politics. Because everyone is concerned with the future, the ability to produce accepted forecasts confers power upon their makers. Another reason for urging more forecasting is the effect on policy. If an individual or collectivity accepts the projected results of a forecast, it becomes the basis for prescriptive action. Humans thus participate consciously to shape their future and to engage in self-fulfilling or self-denying forecasting ("If certain occurrences will happen, we need to undertake the actions to promote, obstruct, or take advantage of them."). In addition to the value accruing to forecasters who are believed and the policy implications of forecasts, certain types of forecasting lead to the expansion of knowledge. If the forecasts have involved articulate calculations or other explicit methods, investigators can presumably use forecasts that prove inadequate to revise their procedures. New estimates of future developments can be made using the revisions and these in turn can subsequently be checked to provide a further round of modifications in the underlying forecasting procedures. Such a cyclical process produces successive approximations that hopefully achieve a gradually improved fit between forecast and subsequent observation. With improved forecasts derived in this

The authors acknowledge the support of the Mershon Center and the Project for Theoretical Politics at the Ohio State University in the preparation of this chapter. They are grateful for the constructive comments of Nazli Choucri on an earlier draft.

fashion should come improvement in the explanatory base that generated them.

Perhaps few proponents of greater forecasting in international relations would state their case in such unqualified terms, but the reasons advanced above appear to capture the core of such advocacy. Notice that all the arguments for more forecasting in international relations assume that someone can eventually determine their accuracy. A forecast that is stated in such a way as to permit its verification against the unfolding future or previously uninvestigated historical events (retrospective forecasts) introduces the problem of forecast validity, which is the subject of this chapter.

The difficulties arise in moving from these simple statements of aspiration to the development of insights and procedures that can be applied in research. At the point of actually validating forecasts a host of philosophical and practical questions arise. What does a forecast represent? Or, put a different way, assuming that a forecast could be validated, what does it mean? How does purpose affect the validation of a forecast? What validation procedures can be employed? What about inconsistencies between the results of forecasts and other means of validating insights or a theory? How can one confidently know (and measure) the future reference system about which a forecast has been made? These questions alert the reader to the conclusions to be found at the end of this chapter. Using forecasts as a validation procedure is much more complex and the results less certain than appears at first glance. Nevertheless, it is an important, if insufficient, operation for improving our knowledge of international relations. For that reason, this chapter seeks to provide some exploration of the issues posed by the questions above and, where possible, to suggest procedures for dealing with them.

II. THEORY AS THE GENERATOR OF FORECASTS

Assume for the moment that by some means a forecast has been validated, by which we mean the state of affairs it asserts as transpiring in a given system has been confirmed as having occurred. One question that remains is what do we know when we have such a validated forecast? In such circumstances, we would know that a particular estimate made at some prior time has been confirmed to some degree by subsequent developments. This confirmation of forecasts can be variously referred to as validation, goodness of fit, verisimilitude, verification, or accuracy. A validated forecast can be used to bestow blame or praise. (Who failed to act upon Senator Keating's warning in the autumn of 1962 that there were missiles in Cuba?) Usually, the accuracy of a single forecast in and of itself is an issue for works of history and biography. Beyond this use of the confirmed relationship between the forecast and actual events, we frequently want to infer something about the means and the source by which the forecast was generated. More specifically, we wish to determine the ability of that source to generate other valid forecasts. Take the following example: "Carl was correct in anticipating the outcome of this week's soccer game, but will his judgment be as good for next week's match?" In this case, the inquiry is about the ability of an individual to make a forecast. Unless he was making an ungrounded guess, the forecaster performed some calculations that formed the basis for his estimate. Thus, one of the fundamental uses of validated forecast is to assess the utility of the calculations by which it was made in order that the calculations can be used again.

As several chapters in this book make clear, numerous ways exist to generate forecasts. When the purpose of one or more forecasts is to determine the utility of an explanatory source for subsequent forecasts and explanations, the components of the forecasting system and their logical relationships to one another must be explicit. Otherwise, what can be inferred about the validity of any future performances of the system will be quite limited.[1] In short, we assert that in order to use forecast validation as a means for inferring the future predictive capability of the source, the source should have the characteristics of a deductive theory. That is, a series of the statements and the logical relationships between them are necessary to derive the forecast. In some instances a given theory may be incomplete in the sense that not all the statements and their connections may be identified, but the closer it comes to approximating the requirements for deductive theory, the greater the value of forecasts as a validating technique.

This theory requirement certainly limits the range of sources that can be evaluated through forecast validation. Nevertheless, the requirement

of a deductive theor
seems appropriate, if
into account the follo

1. Forecasts are us
the source for future

2. The forecast sou
tered to attempt to i
based upon its previc

3. It is necessary t
boundaries beyond w
sharply with respect t

4. The forecast co
system that is suspect
nents that can assume
which in turn may yie

We believe these c
the international rela
the validity of forecas

Before proceeding
to offer some definitio
been using. A *deduc*
set of sentences clos
the set contains any
plied by any other ser
and Stever, 1974, pp
this definition). Furth
are generally asserted
an accurate descriptic

A *forecast* is a state
the state of some w
some other time. Thu
ered for forecasting
the sense that the val
are related to values
points in time.

More precisely, co
reference system con
$x_2, \ldots, x_n$). We war
tences relating at leas
to previous states of
example, these sente
differential equations

$$\frac{dx}{dt} = f_i(x_1$$

The theory of
Richardson illustrates
international relations
Richardson used diffe

of a deductive theory as the source of forecasts seems appropriate, if our validity studies must take into account the following considerations:

1. Forecasts are used to estimate the utility of the source for future forecasts.
2. The forecast source is to be adjusted or altered to attempt to improve subsequent forecasts based upon its previous performance.
3. It is necessary to establish the parameters or boundaries beyond which the source may decline sharply with respect to the accuracy of its forecasts.
4. The forecast concerns a dynamic reference system that is suspected of containing some components that can assume a substantial range of values which in turn may yield quite different outcomes.

We believe these conditions frequently confront the international relations scholar who evaluates the validity of forecasts.

Before proceeding further, it would be desirable to offer some definitions of the basic terms we have been using. A *deductive theory* is stipulated as a set of sentences closed under deduction; that is, the set contains any sentence that is logically implied by any other sentence in the set (see Thorson and Stever, 1974, pp. 15–32 for an explication of this definition). Further, the sentences in a theory are generally asserted to be true; that is, to provide an accurate description of some reference system.

A *forecast* is a statement made at one time about the state of some world or reference system at some other time. Thus, the theories to be considered for forecasting must be dynamic theories in the sense that the values (states) of some variables are related to values of other variables at other points in time.

More precisely, consider a theory about some reference system consisting of state variables $(x_1, x_2, \ldots, x_n)$. We want our theory to contain sentences relating at least some of these state variables to previous states of the system. In physics, for example, these sentences are often expressed in differential equations of the form:

$$\frac{dx}{dt} = f_i(x_1, x_2, \ldots, x_n).$$

The theory of arms races developed by Richardson illustrates a theory of this type in the international relations literature (1960a and 1960b). Richardson used differential equations to relate a nation's level of defense at one time to states of the system at previous times. Forrester's world dynamics simulation offers a second example (1971b). The sentences are in the DYNAMO language and levels of variables at one time are related to levels at previous times. This theory contains statements in the form of difference equations.

In principle, a theory need not be expressed in an artificial language (such as DYNAMO or differential equations) to be a deductive theory. Theories expressed in a natural language, such as English, may also satisfy the above conditions. It might be argued, for example, that Galtung's "rank theory" meets the criteria established above (1964, pp. 95–119). However, a person analyzing most natural language theories (including Galtung's) encounters a difficulty in attempting to unambiguously identify the objects and relations being discussed.

The analysis that follows excludes means of generating forecasts which are not dynamic theories of the kind identified above. Thus, we do not consider trend and cyclical analyses that simply project prior patterns without any antecedent explanations. Nor do we include forecasts generated from the development of speculative or plausible scenarios or Delphi techniques (Morgenstern, *et al.*, 1973). Each of these forecasting techniques lacks to a greater or lesser degree a bounded explanatory system whose component elements can assume different values. Forecasts that are extrapolated from observable trends and apparent cycles in world affairs involve no explanatory mechanism that can be adjusted if the projections fail to correspond to subsequent occurrences. Nor do they contain parameters that permit one to determine whether the conditions of the system at a given time parallel those from which the trend or cycle is derived. Sequential exchanges between panels of experts (as in the Delphi technique) may result in the gradual emergence of consensus around one or several forecasts. But the concurrence may be achieved by using multiple explanations for the same forecast. One explanation may convince some experts, while another persuades other experts to accept the same forecast. Or the explanation may be sufficiently ambiguous so that different authorities are able to attach quite different meanings to its key elements (or substitute their own when these elements are missing). Moreover, consensus

may emerge from group influences processes that have little to do with the merit of the accepted explanations. Lest we be misunderstood, it should be emphatically stated that all these forecasting techniques may have a role in international relations. This chapter, however, concerns the validation of forecasts to improve theory. Evaluation of the validity of the forecasts from such sources as trends, cycles, scenarios, and Delphi techniques, has limited utility for their development.

Now that the class of theories to be discussed has been identified, it is appropriate to return to the concept of "validity" which was briefly defined at the beginning of this section. In discussing a concept such as validity it is important to distinguish between semantic questions of what it means to predicate validity of a forecast, and methodological questions of how it becomes known whether the assertions from a particular theory are, in fact valid. Answers to the methodological question would seem to presume adequate answers to the semantic one. That is, it would be difficult to determine whether a theory is valid without first determining what is meant by validity. Therefore, the first task will be to explicate what will be meant in this chapter when validity is predicted of a theory. Roughly, a theory—a set of logically related sentences in some language—is valid if it does what it purports to do. Thus, as is noted by Forrester (1961) and Hermann (1967, pp. 216–231), the question of validity is inextricably intertwined with the purpose to which, in this case, a forecasting system will be put. A number of possible purposes and criteria of validity appropriate to these purposes will be treated subsequently. With this clarification, we stipulate that a theory, T, is valid with respect to purpose, P, to the extent T achieves P.

Relating validity to purpose is compatible with an extremely pragmatic view of theory evaluation. This compatibility, however, does not require that we adopt such a pragmatic view. One might argue, for example, that the purpose of a scientific theory is to generate (or be capable of generating) "true" sentences (Popper, 1965, pp. 223ff). Thus, the test of validity of a scientific theory is whether the sentences comprising the theory (as well as those logically implied by these sentences) accurately account for and describe features of some reference system. That is to say, for a scientist taking this position to assert that T is a valid theory is equivalent to his asserting that the sentences comprising T accurately account for the operation of the reference system. Note again that this semantic definition of validity does not entail any particular methodological position as to how a particular theory is known to be valid (i.e., known to consist of true sentences). For example, although it might be argued that the goal of science is to construct true theories (i.e., theories whose sentences accurately represent the operations that control the relationship between components of a reference system), yet it could still be argued that it can never be known whether any particular sentence is in fact true.

It will therefore be useful to consider a variety of methodological positions which can be brought to bear on validity questions. Examples of such positions include rationalism, empiricism, and "positive theory." (Naylor and Finger, 1971). The rationalist perspective generally holds that a theory is simply a set of deductions from propositions of unquestioned truth. Thus, no empirical testing is necessary and instead efforts should be spent searching for the basic assumptions from which to generate the theory.

The empiricist response is to refuse to admit any assumptions that cannot be independently verified through controlled observations. The empiricist position—or at least a moderated version of it—is quite evident in the contemporary study of international politics. For instance, Singer (1972, p. 6) argues that the route to explanatory knowledge typically follows a progression: "Just as existential descriptions must precede correlational propositions, so the latter are an essential prerequisite to that explanatory knowledge which is the ultimate goal." Unfortunately, it is not clearly the case that correlational knowledge will lead to explanatory knowledge or that it must precede it. Regression coefficients are properly used to estimate population parameters only when the structure of the theory employed in forecasting is well specified. Data analysis strategies (such as regression analysis) cannot in general reveal the underlying structure of a referent system. This is generally the case whether the systems are analyzed cross-nationally at a point in time or individually as a time series. Thus, it would appear that prior empirical analysis is not a justification for validating a theoretical explanation or forecast of future events in and of itself. We must have a
before any empirica
of validity for our p
that simple extrapol
not scientific foreca
the theoretical expl
trapolation should
which can forecast f

A third perspecti
theory" (Friedman,
1973; and McGowan
tive theorist argues
and empiricist posit
tions ought not be a
utility of a theory de
its assumptions but
which it predicts val
points. Each of the
has been subjected
the positions themse
ent forms to form m
distinct methodologi

There are, howeve
phasis and these have
to suggest a multist
which "each of the a
cal positions is a nec
tion] but that neither
dure for solving the p
Thus, in line with the
sumptions are seen a
and preliminary emp
the less obvious ones
we should subject o
test. However, for a v
possible to test all ass
fore examine the abili
rate predictions abou
the multistage approa
variety of methodolo
points of the ongoing

III. SOME CONSID
THE RELATIONSH
AND FORECAST

Let us take a brief re
in this chapter must
statements concerning
jects at different poin

self. We must have a theoretical structure specified before any empirical analysis provides some sense of validity for our propositions. Popper has argued that simple extrapolations from past to future are not scientific forecasts: "Indeed, unless we have the theoretical explanation for why such an extrapolation should hold, we do not have theory which can forecast future events" (Popper, 1959).

A third perspective might be termed "positive theory" (Friedman, 1953; Riker and Ordeshook, 1973; and McGowan, 1974, pp. 25–44). The positive theorist argues that contrary to the rationalist and empiricist positions, the validity of assumptions ought not be a central question. Instead, the utility of a theory depends not upon the validity of its assumptions but rather upon the accuracy with which it predicts values of variables at other time points. Each of the three positions sketched here has been subjected to considerable criticism and the positions themselves are held in enough different forms to form more of a continuum than three distinct methodological perspectives.

There are, however, distinct differences in emphasis and these have led Naylor and Finger (1971) to suggest a multistage approach to validation in which "each of the above mentioned methodological positions is a necessary procedure for [validation] but that neither of them is a sufficient procedure for solving the problem of validation" (p. 156). Thus, in line with the rationalist position, some assumptions are seen as more "obvious" than others and preliminary empirical work should focus upon the less obvious ones first. To the extent possible we should subject our assumptions to empirical test. However, for a variety of reasons it will not be possible to test all assumptions and we must therefore examine the ability of the model to make accurate predictions about the referent system. Thus, the multistage approach attempts to incorporate a variety of methodological perspectives at various points of the ongoing process of validation.

III. SOME CONSIDERATIONS AFFECTING THE RELATIONSHIP BETWEEN THEORY AND FORECAST

Let us take a brief review. The theories of interest in this chapter must generate forecasts, that is, statements concerning changes in the values of objects at different points in time. We contend that the major question of forecast validity is actually one of using the forecast to assess the validity of the theory that generated the predictions. This does not mean that we can ignore the validity of the predictions themselves. The assertion that under certain conditions a particular pattern of events will occur during some future period of time suggests an obvious criterion for establishing validity of the theory. If the specified conditions transpired, did the projected pattern occur as predicted? The accuracy of forecasts is certainly an essential feature of the validation effort, but a number of issues must be taken into account in evaluating the relationship between a theory and its forecasts.

As we noted in the previous section, the user's purpose should determine whether inferences about the theory from confirmed forecasts are of major importance. Elsewhere some distinctive purposes of simulations (one type of theory) have been described together with their implications for validity (Hermann, 1967). Among the purposes mentioned were (1) the discovery of alternatives, (2) the evaluation of alternative outcomes, (3) prediction, (4) instruction, (5) construction of hypotheses and theory, and (6) exploration of nonexistent universes.

For the present, we need only establish that the user's purpose will make a difference. For example, if the user seeks explanation for why certain macro patterns seem to hold, then the confirmed forecast may be of minimal value in assessing a theory's explanatory adequacy. It is quite possible for a theory involving a number of stochastic processes to yield accurate forecasts about a closed system without providing much insight into why the particular pattern occurs as it does. With respect to the degree of accuracy in forecasting, numerous illustrations come to mind. In deciding whether to sell a particular weapons system to a Persian Gulf country, a United States policy maker may only be concerned with a forecast that qualitatively assesses whether the proposed sale would be stabilizing or destabilizing. No precise quantitative forecast would be required. On the other hand, a theory that estimated the number of ICBM launchers that could be built by the Soviet Union or the United States without detection by the other side would have to have a much higher predictive capacity if it were to be used as the basis for sign-

ing or, not signing, an arms limitation agreement. The users need must determine the precision.

In assessing the degree of precision necessary for the user's purpose, one criterion must be the alternatives available for forecasting. In statistical tests, forecast performance is often compared to change, but that may not be the relevant standard in a particular case.

Another issue we must address concerns the role of probability in the theory. Suppose we have a theory that leads to the following assertion: When nations of the world are ranked according to military and economic capability, the first-ranked nation will initiate war with the second-ranked nation if—and only if—the latter's rate of growth in both military and economic capability relative to the first-ranked nation will lead to a reversal of ranks within 5 years. Such a statement can be contrasted with one that concludes that the first-ranked nation is *more likely* to initiate war against the second if its projected economic and military growth rate will cause it to overtake the first-ranked nation within 5 years. The first statement claims to contain all the conditions that are necessary to produce the projected outcome. The first assertion is that the outcome will occur every time the conditions are met. The second assertion contends only that the specified conditions increase the likelihood of the outcome. Although the examples may seem a bit farfetched, some theories can generate forecasts that are held to be completely determined by the configuration of specified conditions; whereas others are probabilistic theories and provide only projections of the probability associated with various classes of events.[2] When the theory's specified prior conditions are not related in a deterministic fashion to the estimated outcome, a forecasting exercise can provide only limited insight into the theory's degree of validity without consideration of the impact of exogenous variables, such as random disturbances, that operate independently of the system to produce similar outcomes. Moreover, even in the case of the deterministic theory, the lack of congruity between forecast and outcome may lead no further than to recasting the relationship in probabilistic terms.

A deterministic theory yields a set of expected values in some future state but makes no provision for the outcome if the expected values do not occur. It is as if our theory projected the rate of descent of a ball of a certain mass down an inclined plane having an angle that is a certain number of degrees from horizontal, but taking no account of the surface of the plane and the ball, etc. Or, consider the example of theory that projects that a particular rate of economic development in a less developed country will begin, at a given point, to generate a certain amount of capital. These theories neglect what happens if the forecasts are not fulfilled—the amount of friction drastically slows the ball or internal revolution slows capital formation.

If the distribution of outcomes around the projected one involves only gradual deviations, we still might give the theory "high marks" even if slight errors occur. If the distribution of outcomes surrounding the one that is forecasted falls off sharply, then a deterministic theory poses severe problems—particularly if the forecasted outcome is regarded as desirable and those around it appear undesirable. Therefore, although forecasts of a deterministic theory may more readily be tested for their validity, inaccuracies may be more difficult to interpret (i.e., how far off is the actual outcome?) and pose serious difficulties for some purposes (e.g., policy analysis).

Actual international political systems have a counterpart problem to the deterministic-probabilistic characteristic of theories. We must consider the actual distribution of the forecasted events in international relations. Are the occurrences conceptualized as unique and noncurrent or are they defined so as to be repeated regularly? Examples of the former include the death of Mao or the acquisition of nuclear weapons by Japan. The frequency of changes in a country's political leadership, illustrations of recurrent phenomena, or the rate of diffusion of a technology are illustrations of the latter. If the phenomena that are the subject of the theory reoccur in the reference system, we need to take into account the frequency of their appearance. Are they frequent occurrences—such as diplomatic exchanges or trade negotiations—or relatively less frequent—such as interstate wars or global economic depressions? Suppose that a theory's forecast of the probability of the outbreak of war under certain conditions is .75 and in subsequent actuality the conditions are fulfilled but no war occurs. Over a series of such forecasts we could establish whether the forecasts correspond to events three-fourths of the time, provided that the class of predicted events

occurred with suffi
the set of condition
we would have a sit
in weather forecasts
ity of rain in the r
precisely, the prob
under conditions su
prevail in this local
Unfortunately, ther
ternational relations
quency with which
earth. Thus, we hav
can predict a patte
occur in the real wo
assess with confiden

One thoughtful c
author in his previo
to consider that an
criteria for validatin
misinterpretation of
world"—rather tha
(Powell, 1973). The
lem in the inferent
casts and theory. W
tween forecasts and
might ask whether i
us call it theory X—
another theory—des
serve and interpret t
astronomer calculat
movement of other b
a previously undetec
able at a certain poi
is the astronomer's
wrong or should we
or the laws of physi
jects in space relativ
forecasts a certain
growth that is not
economic activity as
reexamine the simul
economic performanc

Certainly, a comm
sider all such avenue
pear to be at variance
evant reference syste
the investigator to de
ing which explanatio
forecast he or she s
theory of optics been
in other tests? Does

occurred with sufficient regularity together with the set of conditions specified in the theory. Then we would have a situation comparable to that used in weather forecasts of precipitation: "The probability of rain in the next 24 hours is .80—or more precisely, the probability of precipitation is .80 under conditions such as those that are expected to prevail in this locality during the next 24 hours." Unfortunately, there are numerous events in international relations that do not occur with the frequency with which rain falls on many parts of the earth. Thus, we have a situation in which a theory can predict a pattern of occurrences that do not occur in the real world with sufficient regularity to assess with confidence.

One thoughtful critic has charged that the first author in his previous writing on the subject failed to consider that an error in forecasting (or other criteria for validating a model) can result from a misinterpretation of the reference system—or "real world"—rather than from an inadequate model (Powell, 1973). The charge highlights another problem in the inferential relationship between forecasts and theory. When an incongruity exists between forecasts and subsequent developments, one might ask whether it results from the theory—let us call it theory X—that generated the forecasts or another theory—designated theory Y—used to observe and interpret the reference system? When an astronomer calculates from deflections in the movement of other bodies in our solar system that a previously undetected planet should be observable at a certain point in space and none is found, is the astronomer's theory of the missing planet wrong or should we reexamine the theory of optics or the laws of physics used for locating other objects in space relative to the earth? If a simulation forecasts a certain pattern of national economic growth that is not substantiated in subsequent economic activity as measured by the GNP, do we reexamine the simulation or the indicator of actual economic performance?

Certainly, a committed scientist ought to consider all such avenues in cases of forecasts that appear to be at variance with occurrences in the relevant reference system. It ought to be possible for the investigator to develop a strategy for determining which explanation for the lack of a confirmed forecast he or she should pursue first. (Has the theory of optics been substantiated independently in other tests? Does the present test use GNP in ways the measure has not previously been used?) With respect to simulations, it is often concluded that inaccurate forecasts are indicative of inadequate theory as represented in the simulation. Perhaps such inferences are too easy. Our conceptualizations and observation techniques in international relations have seldom been confirmed in a systematic fashion. In a given area of international relations there may be no definition of the key concepts, no explicit statement of assumptions, and very unreliable measures of observation. Under such circumstances, the scholar must be acutely sensitive to the possibility that the means for verifying the forecasts require careful examination. This point will be considered further when measurement problems are discussed in the next section.

Another type of problem arises in instances having substantial goodness of fit between forecast and reference system events. How confidently can we infer from such verisimilitude in the forecasts to the theory assumed to have accounted for the observed developments? The possibility exists that the correspondence of events and forecasts results from spurious correlation, coincidence, or overdetermined events. The appearance of a substantial goodness of fit that actually resulted from fortuity should be eliminated by repeated forecasting attempts that would reveal the coincidence as random error. Repeated tests should also identify those situations that are overdetermined—that is, outcomes produced by any of several different factors and all of which happen to be present in a given instance. Across a variety of forecast occasions, some of the relevant conditions may not occur, and those accounted for in the theory will be responsible for the observed result. Somewhat more troublesome is the systematic error in the form of a spurious correlation. Although repeated forecast efforts may reveal the presence of this problem, one also can put the theory in an operational form—or simulation—and conduct sensitivity tests to determine the effects of individual components on the outcome when other elements are held constant.

The use of sensitivity testing to check for spurious correlations introduces a point applicable to all the issues discussed in this section. In order to clarify problems that can affect the assumed relationship between a forecast and the theory that generated it, we must examine the theory directly.

For spurious correlations we want to conduct sensitivity tests on the theory, perhaps, as represented in a simulation model. To determine the implications for forecasting of the user's purpose, we need to examine the theory for its correspondence with such purposes. If we have a deterministic theory, we need to identify with special care the variables not contained in the theory that could alter the forecast. Should the theory predict rare events in the reference system, we need to establish estimates of our confidence in the theory independently of its forecasts of those infrequent occurrences. (We will return to this point in the discussion of plausibility in the next section.) Again, in deciding between errors in theories that generate forecasts and errors in theories involved in assessing the actual occurrences in international politics, we must move outside the forecasts themselves.

In short, issues that can affect our inferences about theory which are made from confirmed forecasts require us to deal directly with the source. This observation is one reason why we contend that if validity concerns us more than the forecast itself, then the source of the forecast must be an explicit theory. Unless the source of the forecast reveals its components and their relationships, resolution of the issues discussed in this section often becomes impossible.

IV. VALIDATING THE FORECASTS

To say that a forecast is "valid" certainly involves making claims about the correspondence between the events asserted to obtain in the forecast and the events that do, "in fact," obtain in the referent system. However, forecast validity should not be viewed only in terms of the correspondence between forecasted events and observations made of the world. At the time a forecast is made, it describes future events.[3] Whether this description is "accurate" or not depends upon events that have not yet occurred. We still may want to make statements about the validity of the forecast. However, the precise empirical testing of forecasts is often quite expensive in time and money and we ought first to satisfy ourselves that such an effort is justified. Therefore, it will be useful first to consider several preliminary tests to which theories used in forecasting might be subjected, prior to explicitly confronting the forecast with observation-based data.

First, it seems reasonable to establish that the forecasts be plausible. By plausible we mean that the forecasted events do not grossly contradict present understanding of the way the referent system behaves. As an illustration consider the story—perhaps apocryphal—of the response of one of the developers of an American quarterly econometric model who was asked what he would do if his model forecast a 15 percent unemployment rate. He answered that he would ignore the model because no US government would permit unemployment to reach 15 percent. In other words, to him such a forecast would be highly implausible. One means of evaluating the plausibility of forecasts is to consult with people who deal with the particular domain about which the forecasts are made. Policy planners, for example, often are able to make informal judgments regarding the probable consequences of actions. The evaluation of these experts provides information useful in constraining the class of plausible forecasts.

Another method of testing plausibility is related to the point made earlier about sensitivity testing. Occasionally a theory that generates plausible forecasts when the values of variables are held to expected or previous levels yields absurd results if certain values exceed "normal" levels. For example, some education planners argued for a theory that forecast exponential enrollment growth in higher education. Predictions from the theory seemed to fit the data very well until about 1969. However, the model predicted exceedingly larger student enrollments for more extreme values of time. For the late twenty-first century, the number of US college students was forecast to exceed the total predicted population of the United States. Thus, we would want to be wary in using such a theory in making very long-range forecasts. Systems "stressing" of this kind is frequently ignored because the theory makes quite plausible predictions in shorter time frames or for more normal ranges of events. However, even very simple sensitivity tests such as that outlined above, may reveal that much of the process about which a theory forecasts is not yet understood.

In considering the empirical aspects of validation, one of the important questions concerns how observational data should be employed to accept or reject the propositions in the theory. Resolving such problems is one of the tasks of inferential statistics. The logic underlying the use of statistics

in validating forecasts
tain propositions in t
are related in specif
gether with measure
variables are used to
ables at other times
tions $p_1, p_2, p_3, \ldots,$
$f_2, \ldots, f_m$. From log
$f_m) \rightarrow \sim(p_1, p_2, \ldots,$
cast statements are f
the theoretical propo
ing these forecasts. A
of the procedure is r
measurement and ob
tion of the strategy d

The problem is es
dynamic and complex
in international politi
the referent system i
assuming we want to t
particular independen
ance accounting techn
is not generally an a
(1973) demonstrates t
in the disturbance te
leads to a serious ove
independent variables
divided into two partic
when there are no lag
overestimation effects
tion of the regression
the importance of the
the second case where
the analysis, not only
but the actual level of
influenced in such a w
variables' importance i
ables' importance is in
decreases can be of a r
cent.

Although the prob
largely from assumed
arise when we conside
ables and complex
nonstationary feedback
many interesting aspe
In such cases, a possib
the theory into "subt
subtheory independen
there is no guarantee
work. For example,

in validating forecasts is fairly straightforward. Certain propositions in the theory state that variables are related in specific ways. These relations together with measures of observed values of these variables are used to forecast values of these variables at other times. More specifically, propositions $p_1, p_2, p_3, \ldots, p_n$ together imply forecasts $f_1, f_2, \ldots, f_m$. From logic we know that $\sim(f_1, f_2, \ldots, f_m) \rightarrow \sim(p_1, p_2, \ldots, p_n)$. Thus, if some of the forecast statements are false then so must be some of the theoretical propositions employed in generating these forecasts. Although the underlying logic of the procedure is relatively simple, problems in measurement and observation make implementation of the strategy difficult.

The problem is especially troublesome for the dynamic and complex referent systems of interest in international politics. For example, given that the referent system is parameterized by time and assuming we want to test the relative importance of particular independent variables using normal variance accounting techniques, ordinary least squares is not generally an appropriate technique. Hibbs (1973) demonstrates that if auto correlation occurs in the disturbance terms, ordinary least squares leads to a serious overestimation of the impact of independent variables. This impact can be subdivided into two particular classes. In the first case, when there are no lag variables in the analysis, the overestimation effects do not influence the prediction of the regression coefficient but they do affect the importance of the t test or the multiple R^2. In the second case where lag variables are included in the analysis, not only are the above effects noticed, but the actual level of the regression coefficients is influenced in such a way that usually the nonlagged variables' importance is decreased and the lag variables' importance is increased. These increases and decreases can be of a magnitude of 300 to 400 percent.

Although the problems sketched above stem largely from assumed dynamics, further problems arise when we consider the large number of variables and complex relations (e.g., nonlinear, nonstationary feedback, etc.) that probably occur in many interesting aspects of international politics. In such cases, a possible strategy is to decompose the theory into "subtheories" and evaluate each subtheory independently. However, in general, there is no guarantee that such an approach will work. For example, Ando, Fisher, and Simon (1963) have demonstrated that only when dealing with linear relations and only when the variance to be accounted for is explainable by the variables in each decomposed subset will such a decomposition strategy work. They proceed to show that it is more frequently the case that the subsystems are only partially decomposable (most, but not all, variance is explainable by variables within the subset). In such cases the subsystem can be treated independently only over short periods of time. Over long periods of time, interactions between subsystems become dominant. Thus, in longer range forecasting it is generally an unwise strategy to attempt to break a theory into more manageable subsets having fewer variables. This conclusion is similar to that of George who suggests that, at least for policy making, theories with more variables may have greater utility (1971, p. xvi).

As was mentioned above, when confronting forecasts with observations to determine empirical validity, it is necessary to make assumptions regarding the procedure for measuring the referent system. These assumptions are necessary because employing any of the statistical strategies outlined previously requires that predicted values of the measures be compared to actual values of the measures. If the actual value is not interpretable in the same terms for both the referent system and the theory used to make the forecast, then any statistical comparisons will be suspect. In other words, it is important to establish the validity of the measures used to assess the accuracy of forecasts.

Not only must measures be valid, but as was mentioned earlier, they must also be accurate enough for the purposes to which they will be put. The accuracy of measures is especially important when doing forecasts that use present values of variables to project future values. As an example, consider a theory that proposes (Fucks, 1965, and Morgenstern, *et al.*) that the change in the power (M) of a nation at one time is some constant times its present power (M_t):

$$M = \rho M_t.$$

Letting M_o be the initial value, M_t for any t can be computed by:

$$M_t = M_o e^{\rho t}.$$

Suppose ρ is measured to an accuracy of ± 3 percent. Numerically, suppose the estimated value of ρ is 2.0. Then $\rho = 2.0 \pm .06$. By the time $t = 10$,

the 3 percent error in ρ will have compounded to a more than 25 percent range in the predicted value of M_{10}.

Similar sorts of things happen when using the more standard one step linear model:

$$X_{t+1} = aX_t + \epsilon$$

where ϵ is a disturbance term with variance σ^2 and mean 0.

For example purposes, suppose $a = 1.01$, the expected value of $x_1 = 7$ and $\sigma^2 = 1$. At time 1 then the expected value of X is 1. If we set a tolerance region of three standard deviations, the observed value should be 1.0 ± 7.9. Using the linear model, the expected value of X_{25} is 1.3. However the variance of X_{25} is 41.75. Using the same three standard deviation tolerance region, the observed value should be 1.3 ± 19.38. Indeed, in general if the system being theorized about is not stable ($a < 1.0$), the farther out in time one projects, the greater will be the variance. In the example above, $\text{var}(t_{50}) = 100.7$. $E(X_{50})$ is 1.6. The three standard deviation tolerance region is 1.6 ± 30.1. Very small measurement errors are often greatly compounded in long range forecasts.

V. SUMMARY

This chapter has attempted to develop a series of observations about the validation of forecasts. They are summarized in the following numbered points.

1. Often our interest in confirming forecasts in international relations is to facilitate our judgment about the source of the forecast. For example, we may wish to evaluate the source in order to establish our confidence in its ability to make future forecasts.

2. A number of factors can affect the relationship between the forecast and the source that generated it, leading to incorrect inferences about the source. Among these problems are the effects of the user's purpose, whether the generating theory is deterministic or probabilistic, the frequency of occurrence of the forecasted events in international relations, the adequacy of the theories for measurement and interpretation of the reference system, and the confirmation of the forecast by observations other than those used in making the projection.

3. These obstructions to reasonable inference about the source can often be assessed if we can examine the source of the forecast in various ways. Such independent testing of the source is possible only if the components of the theory and their relationship are known and precisely defined. For this reason, we contend that the source of the forecast must be a deductive theory. Other sources of forecasts in international relations may produce valid forecasts and play a vital role. But problems will be encountered with them if we try to assess systematically their potential for subsequent forecasting efforts.

4. The task remains of determining the goodness of fit between the theory's projection of the future value of certain variables and the subsequent unfolding of actual occurrences. Before beginning empirical testing, some efforts should be made to determine if the forecast is within contextual constraints, that is, whether it is plausible. Empirical testing with statistical techniques can follow, but the investigator should be mindful of several factors—such as consistency in level of analysis—that can influence results.

NOTES

1. A major difficulty with much contemporary international relations forecasting is that the calculations, the conceptualizations, the mental images, the models in the mind—whatever we label the cognitive processes used to derive a forecast—often remain implicit and unarticulated. Kaplan has summarized the resulting problem in a delightful way. "Too often the hypotheses with which we work are at home only in the twilight regions of the mind, where their wavering outlines blend into a shadowy background . . . . Forced into the open, our ideas may flutter helplessly; but at least we can see what bloodless creatures they are" (Kaplan, 1964, pp. 268–269).
2. The distinction between the projected outcomes from probabilistic as compared to deterministic theories overlaps somewhat with Choucri's distinction between predictions and forecasts. We maintain, however, that a deterministic theory could still produce a forecast in Choucri's sense of the term. See her discussion in Chapter 1.
3. In the case of retrospective forecasting the events already have transpired but must be unknown to the forecaster. Hence for the purposes of this decision, they can be treated in the same manner as events that have not yet occurred.