CRITIQUE AND COMMENT

VALIDATION PROBLEMS IN GAMES AND SIMULATIONS WITH SPECIAL REFERENCE TO MODELS OF INTERNATIONAL POLITICS[1]

by Charles F. Hermann

Princeton University

To what extent are simulations and games valid representations of behavioral systems? Although this frequently asked question is vital, it concerns only one aspect of the larger issue of operating model validation. This article describes how the purpose of the experimenter, the type of validity criteria, and the use or nonuse of human participants each influences validation. Many of the illustrations are drawn from games and simulations of international politics, but the observations are intended to be applicable to similar activities underway throughout the behavioral sciences.

TECHNIQUES of gaming and simulation are receiving widespread attention throughout the behavioral sciences. We can observe, for example, the proliferation of these techniques in political science. Games and simulations currently exist to explore such topics as local politics, legislative processes, community and national election campaigns, public opinion formation, political bargaining, politics in developing countries, American national government, international politics, and disarmament. Indeed, in areas like international politics or election processes, a potential user can choose from four or five different models. Many of these techniques are being used for instruction, others for research, and some as adjuncts to policy making.

A simulation or game is the partial representation of some independent system. Usually we are interested in simulation as a means for increasing our understanding of the system it is intended to copy. Therefore, the representativeness of a simulation or game becomes extremely important in assessing its value. The process of determining how well one system replicates properties of some other system is called validation. In experimental research, validity is the goodness of fit or the correspondence between phenomena produced by two sets of properties. Thus, a test designed to measure dogmatism is said to be valid if the test scores of individuals correspond highly with the extent of their participation in activities stipulated as dogmatic. In the present analysis, however, validation will be defined more broadly as any comparison between the representation of a system and specified criteria.

As a device for the representation of an independent system, the simulation-gaming technique is part of the generic class called models. Thus, it is likely that some issues of simulation or game validity are applicable to other types of models. The distinctive property of a simulation (or game) which separates it from a verbal, pictorial, or mathematical representation of a system is the former's ability to evolve through time. As the related components of a simulation interact with one another and assume different values, the model can take on states

[1] An earlier version of this paper was presented at the American Political Science Association, Chicago, Illinois, September, 1964. Some ideas incorporated in the present version were gained by the author's participation in a Conference on Problems of Validation in Simulation held at Northwestern University in February, 1965 under the joint sponsorship of the Joint War Games Agency and the Advanced Research Projects Agency. A special indebtedness is due to Harold Guetzkow who directed that conference and Abraham Kaplan who commented on the paper in draft form. During the preparation of this article, the author was supported by the Center of International Studies at Princeton University.

vastly different from the one that existed initially. Not only is there a "dynamic quality" (Davis, 1966) to such models, but they also are capable of producing events that an observer could not readily derive from the initial conditions of the simulation. The emergence of unforeseen events that occur as this type of model operates through time have led to the characterization of games and simulations as "operating models" (Brody, 1963; Guetzkow, 1959).

Although they are operating models, simulations and games—like other models—are always a simplification of their reference system. Some features are excluded. Those elements incorporated in the model are often represented so as to reduce the number and complexity of actual components and relationships. Techniques for simplification include the following: compressing a number of different properties into a single, prototypic feature (for example, representing all organizations which influence an individual's opinion by one group); replacing one property for another (such as representing the production of goods and services by chips or by symbols in a computer program); or introducing a single probabilistic function (for example, representing the complex of determinants affecting success or failure in a search and development program as a specific probability of a payoff).

The reduction of the observable universe, necessary for the construction of a simulation, helps to identify the operation of major determinants in the system. Parsimony is a significant contributor to a model's value as a device for prediction and explanation. On the other hand, simplification and abstraction of processes and structures assumed to exist in the referent system reduce the credibility of claims that correspondence exists between an operating model and its intended referent. Simplification in model building increases the uncertainty of a simulation's "representativeness," and thus adds to the necessity for establishing validity.

The contention of this article is that simulation or game validity is not a singular problem. Instead, we may more accurately refer to multiple validity issues. In the present context, three of these problems in the validation of operating models will be discussed. First, the validity of an operating model is affected by the *purpose* or use for which the game or simulation is constructed. What may be a relatively valid operating model for one objective may be strikingly unsatisfactory for another. Second, model validation can be expected to vary according to the type of validity *criteria* employed. Third, the validation issues will be significantly altered depending on whether *human participants* are introduced into the model. Although the author's own experience leads to illustrations of these problems drawn primarily from international politics, the observations presumably are applicable to simulation and gaming techniques throughout the behavioral sciences. Occasional examples will be offered from this wider range of studies. Moreover, some of the comments may be salient to verbal and mathematical models in which the problems of validation appear equally essential and equally neglected.

EFFECTS OF PURPOSE ON VALIDITY

We usually assume that the purpose for using an operating model is to create an A (a game or a simulation) in a way such that it will reproduce as accurately as possible aspects of B (a selected reference system). This objective is frequently found in simulation and game activity, particularly when the purpose is to explain or predict the behavior of B. Not all games or simulations, however, have that purpose. When the primary objective for an operating model is not to replicate aspects of some system, then the model's validity is affected. Outlined below are other goals for gaming or simulation and their implications for validation. This list, however, does not exhaust all possible objectives.[2]

Alternatives and their consequences

Some operating models are designed to generate alternative courses of action and to stimulate consideration of the possible con-

[2] For a fuller discussion of the role of purpose in simulation and game validity, see Kress (1965) and Crow and Noel (1965).

sequences associated with pursuing a particular alternative. Operating models with this purpose are used by policy makers or those who wish to explore policy problems. With these games and simulations investigators hope to uncover aspects of policy management that might not otherwise be revealed. For example, 76.7 percent of 73 government and academic participants in political-military games at MIT indicated on a questionnaire that one of the values of political gaming was to increase the number of alternatives that a policy maker would perceive in a similar (not identical) real situation (Barringer and Whaley, 1965). A statement of the gaming objectives of the Joint War Games Agency (Office of the Joint Chiefs of Staff) reflects their interest in using the technique to explore interpersonal and interagency relations in handling various hypothetical alternatives:

> They [political-military games] are not intended to be predictive. They are played by top level officials from the White House, State, and Defense Departments, and the Services for the following benefits: . . . build inter-agency and interechelon rapport; point up weak spots in coordination, etc.; provide "feel" for Cold War "bargaining," negotiation, and escalation processes; broad overview for specialists; cross-fertilize ideas between agencies (Giffin, 1965, p. 70).

In operating models designed for this purpose, validation does not mean the "correct" replication of the reference system. The validity requirement is that the game or simulation aid its users in such ways as to detect useful alternative means of handling a problem, need for more detailed planning, or requirements for coordination.

Relative predictive ability

When game or simulation developers have this purpose they are interested in the ability of an operating model to predict certain outcomes as compared to the projections of other methods of prediction. If the issue is the outcome of labor-management negotiations, then a simulation of negotiation processes is compared with the predictions of mediators, the press, and so on. If the task is to predict an election outcome, the projections of a campaign simulation are compared to those of polls (Pool, Abelson, and Popkin, 1965). Operating models are usually assigned this purpose in situations in which forecasts must be made. Thus, for example, computer simulations of weather forecasting have been compared with other methods (Thompson, 1961). Even the best method of prediction may be imperfect, but because predictions are required, it is used. In this instance, judgments about validity are not drawn exclusively from the reference system, but also from the alternative means of making projections.

Instruction

Many games and simulations are designed for teaching or training. A validity strategy can be influenced by the instructional purpose in several ways. No operating model should transmit systematic misperceptions or incorrect images of the reference system involved in the instructional program. For this reason some models acceptable for various research activities may be unsatisfactory for teaching. For example, consider a complex operation in the reference system which is simplified in the operating model by a stochastic process. If the probabilities assigned to the various alternatives are correct, the model will display the outcomes accurately. Yet the student has not been given information about how these outcomes are actually obtained. Should he leave the game or simulation believing the outcome is a function of chance, he may be badly misinformed. On the other hand, an operating model which does not necessarily correspond accurately to its intended reference system may still have vital educational attributes if the lack of correspondence is understood by the students. Under these conditions, a game may provide the student with greater empathy for those persons who operate in the reference system than he might otherwise have. Moreover, the game or simulation may enhance motivation of students or trainees in the subject. Use of the Carnegie Tech Management Game as an educational device for graduate students illustrates how instructional objectives may conflict with careful representation of an actual business firm. The game staff has discovered "that presidents, executive vice presidents, and

men in charge of marketing and
learn more than men in other pc
In other words, some game po
regarded as routine and uncl
they offer the fewest opportunit
or to take actions that will aff
tunes of the team" (Cohen, I
and Winters, 1964, pp. 263–264).
is faced with a choice between
more completely the variety o
firm, or a more simplified role st
maximizes the learning experi
students. In summary, the vali
have shifted from the observal
to the effects on the cognitive a
systems of those individuals
operating model is intended to i

Hypothesis and theory construc

In developing a game or sim
designer is required to be explic
nature and relationships betwe
in the operating model and th
parts in the observable univers
specify the conditions which c
tionship to vary. In constructin
ing model a connection betwee
unrelated findings may be dis
ternatively, a specific gap in kn
be pinpointed and hypotheses
the model may be advanced to
explanation. Thus the authors
tion of national policy format
their current objective as the "g
hypotheses and explication of
gested by observations made
preparation and conduct of the e
(Boguslaw, Davis, and Glick, 1
Once the game or simulatio
its contribution to hypothesis
building can be continued by
the completeness of the theory re
the model. A simulation current
velopment (Pool and Kessler,
basic principles of individual
explain international behavio
searchers recognize that the
their model account for only a s
international affairs. But th
simplified model to improve th
world politics. "The best way
what part, if any, individual

men in charge of marketing and production learn more than men in other positions do." In other words, some game positions "are regarded as routine and unchallenging—they offer the fewest opportunities to learn or to take actions that will affect the fortunes of the team" (Cohen, Dill, Kuehn, and Winters, 1964, pp. 263–264). The faculty is faced with a choice between representing more completely the variety of roles in a firm, or a more simplified role structure that maximizes the learning experience for all students. In summary, the validity criteria have shifted from the observable universe to the effects on the cognitive and affective systems of those individuals whom the operating model is intended to instruct.

Hypothesis and theory construction

In developing a game or simulation, the designer is required to be explicit about the nature and relationships between the units in the operating model and their counterparts in the observable universe. He must specify the conditions which cause a relationship to vary. In constructing an operating model a connection between previously unrelated findings may be discovered. Alternatively, a specific gap in knowledge may be pinpointed and hypotheses required by the model may be advanced to provide an explanation. Thus the authors of a simulation of national policy formation describe their current objective as the "generation of hypotheses and explication of ideas suggested by observations made during the preparation and conduct of the experiments" (Boguslaw, Davis, and Glick, 1966, p. 43).

Once the game or simulation is created, its contribution to hypothesis and theory building can be continued by determining the completeness of the theory represented in the model. A simulation currently under development (Pool and Kessler, 1965) uses basic principles of individual behavior to explain international behavior. The researchers recognize that the variables in their model account for only a small part of international affairs. But they use the simplified model to improve their theory of world politics. "The best way to ascertain what part, if any, individual psychology plays in the determination of political behavior is to postulate the truth of the extreme proposition and then see what conclusions it leads us to" (p. 32). In this example, comparisons with criteria from the observable universe will still be required. Like other such enterprises, however, the final validity criteria are in terms of the heuristic payoff from the simulation for hypothesis and theory building.

Nonexistent universes

Throughout this essay we use the term "observable universe" interchangeably with "reference system" to refer to that which a game or simulation represents. The fact must be recognized, however, that some simulations and games are concerned with nonexistent universes rather than observable ones. In other words, some operating models are designed for the purpose of exploring "what-would-have-happened-if" worlds or conditions of the future. Brody (1963) conducted a series of simulations to investigate the possible systemic effects of nuclear proliferation on a bipolar international political system. Similarly, Kahn (1962, p. 174–175) has used the gaming technique to explore various situations involving nuclear weapons on the grounds "that if we are to understand the problems of national security and international order we must be sure to analyze improbable and terrible situations." For the inquiry into nonexistent systems, establishing appropriate validity criteria is difficult. Logical consistency and reliability may be relevant. Insofar as the nonexistent system under investigation is a variant of an existing one, then the approaches which derive criteria from an observable referent (described in the next section) may prove to be of some value. The applicability of these approaches would appear to decrease as the number of properties in the hypothetical system which differ from any in the observable universe increases.

Two observations can be drawn from this discussion of the various objectives of games and simulations. The first comment is that the validation of an operating model cannot be separated from the purpose for which it is designed and used. As two simulation researchers have recently concluded, "... accuracy and precision are no longer the sole considerations in evaluating a method, ... a method 'valid' for some purposes may not be valid for others" (Crow and Noel, 1965, p. 25). The second observation somewhat mediates the first. For the most part the various purposes for conducting games and simulations do not negate the need for criteria we can use to estimate the degree of fidelity with which one system (the operating model) reproduces aspects of another (the reference system). Given some purposes for using games and simulations (such as exploring nonexistent universes), finding appropriate criteria in the referent system is quite difficult. With other objectives, the value of the operating model may remain even if the fit between the model and various criteria representing the observable universe is poor (as in theory building).

Nevertheless, with models designed for purposes such as those we have been describing, it is pertinent to ask such questions as: Are the alternatives developed in this model likely to be comparable to those we will encounter in the actual system? How accurate are the predictions we are obtaining from the best available projection technique? Does the instructional game or simulation bear a close resemblance to selected aspects of the observable universe under examination? Is a theory-building model a reasonable representation of the phenomena which we are ultimately interested in explaining? What correspondence is there between the model of a hypothetical system and aspects of the observable universe? The point is not that these are irrelevant validity questions, but rather that they are not the only validity questions.

TYPES OF VALIDITY CRITERIA

We have been considering the impact of an operating model's intended purpose on validation. A more complete treatment of simulation and game validity would examine the validity criteria associated with each potential objective. In this limited exploration, however, we will confine our attention to criteria which are applicable when the intended purpose for a game (or simulation) requires assessment of the degree of correspondence between the game and aspects of the reference system. The goodness of fit example is instructive for several reasons. First, as we have noted, the fidelity with which a model reproduces aspects of reality is relevant to simulations and games with widely diverse purposes. Second, it readily illustrates a number of the issues involved in establishing validity criteria. Nevertheless, in the following discussion it should be remembered that the use which is made of a particular operating model will introduce types of validity criteria other than those designed to determine the degree of correspondence between the model and the reference system.

To determine the extent of representativeness it is necessary to establish what aspects of the actual behavioral system should be selected as validity standards or criteria. The five approaches described below identify criteria that parallel properties found in most operating models. In all but the first approach we are asking to what features of the observable universe can we extrapolate or generalize operations occurring in a game or simulation.

Internal validity

Suppose the same game is executed a number of times, each trial beginning with identical parameter settings and the same initial values. Any exogenous inputs introduced during the course of the game are held constant across all trials or runs. The unexplained variance between these intended replications would provide a measure of reliability or what Campbell (1957) calls "internal validity." When the structured simulation properties are held constant, the smaller the between-run variance, the greater the internal validity is assumed to be. If the observed results of an operating model can be attributed to extraneous factors rather than the specified relationships in the simulation, then its internal validity is low.

The nature of internal validity is illustrated by the study of nuclear proliferation conducted in a series of runs with the Inter-Nation Simulation (Brody, 1963). The ex-

) requires assessment of the de-
rrespondence between the game
s of the reference system. The
of fit example is instructive for
sons. First, as we have noted, the
th which a model reproduces as-
lity is relevant to simulations and
widely diverse purposes. Second,
llustrates a number of the issues
n establishing validity criteria.
ss, in the following discussion it
remembered that the use which is
particular operating model will
types of validity criteria other
designed to determine the degree
ondence between the model and
ce system.
rmine the extent of representa-
is necessary to establish what
the actual behavioral system
selected as validity standards or
he five approaches described be-
y criteria that parallel properties
ost operating models. In all but
pproach we are asking to what
the observable universe can we
or generalize operations oc-
game or simulation.

lidity

the same game is executed a
times, each trial beginning with
arameter settings and the same
es. Any exogenous inputs intro-
ng the course of the game are
nt across all trials or runs. The
l variance between these intended
would provide a measure of
or what Campbell (1957) calls
validity." When the structured
properties are held constant, the
e between-run variance, the
internal validity is assumed to
observed results of an operating
be attributed to extraneous fac-
than the specified relationships
lation, then its internal validity

re of internal validity is illus-
he study of nuclear proliferation
n a series of runs with the Inter-
ulation (Brody, 1963). The ex-

periment was repeated 17 times. Each simulated international system began with the same initial conditions—the same seven nations, national resources, alliance affiliations, government offices, and so on. So far as the programmed aspects of this man-machine simulation were concerned, the initial values of all the parameters in every replication were identical. Although the participants differed in each run, efforts were made to distribute them among the runs so as to reduce the between-run variance of psychosocial characteristics judged by the experimenters to be critical. Nevertheless in three, and only three, of the runs war occurred between the two opposing superpowers before nuclear weapons were dispersed to other nations (Caspary, 1962). The outbreak of war cannot be explained by the endogenous properties of the simulation. Some as yet unidentified external factor or factors operated differently in three of the runs to alter the outcome, and thereby reduced the internal validity.[3] With regard to operating models, the critical requirement for reliability or internal validity is that variations be accounted for by identifiable relationships within the game or simulation. In internal validity, then, the criterion is the replicated simulation.

Face validity

Face validity is a surface or initial impression of a simulation or game's realism. Probably no approach to model validity is reported more frequently than the subjective estimates of experimenters, observers, or human participants as to the correspondence between the model's operation and their perception of the actual phenomena which the game or simulation represent. In describing their Carnegie Tech Management Game, the authors (Cohen, Cyert, Dill, Kuehn, Miller, Van Wormer, and Winters, 1962) comment: "After a period of development and experimental trials we now have a game which we feel has achieved a kind of complexity and realism desired" (p. 105). A simulated air defense center was also reported to have achieved considerable realism. "Each of the four crews gradually came to behave as if it were in a real-life situation. . . . During enemy attacks, the noise level in the station rose, men came to their feet, and the excitement was obvious. . . . On one occasion, an officer slipped while stepping off a dais and broke his leg. We were not aware of this event for some ten minutes because there was no perturbation in the crew's activity during the attack in progress" (Chapman, Kennedy, Newell, and Bill, 1962, p. 185).

Face validity can be a significant part of a validity strategy. A quick impression that "things don't seem right" may be the only validity check possible during the actual operation of a game or simulation. Such validity judgments and their evaluation may also be part of the learning experience provided by operating models designed for instructional purposes. Furthermore, at least one method for assessing the credibility of simulation findings has assigned a substantial role to the experimenter's intuition about a model's validity. In their extension of Bayesian subjective probability concepts to the problems of validity, Rice and Smith (1964) require the model builder to make subjective probability estimates of the reliability and relevance of his simulation.

Although face validity has value in the early stages of model building or for quick checks during actual operation, its severe limitations should be recognized. Sometimes the experimenter will not know what behaviors are "realistic" because of his limited experience observing the actual phenomena. Participants can become interested and highly motivated in an incorrect representation of the desired environment. If the simulation involves the substitution of one property for another, some features may appear quite unreal and yet replicate the performance of the reference system for which the simulation was designed. The acceptance of face validity as a rough, first approximation might be improved if the simulator explicitly stated in advance what observations would constitute indications

[3] It should be noted in regard to other variables that the same series of simulation runs displayed considerable internal validity. For example, Brody (1963) found no significant between-run variance in the amount of hostility displayed within and between alliances.

that an aspect of the observable universe had been successfully captured. In summary, face validity in its usual form suffers from the lack of explicit validity criteria.

Variable-parameter validity

Another approach to the validation of operating models involves comparisons of the simulation's variables and parameters with their assumed counterparts in the observable universe. Kelman (1965) is among the social scientists who have stressed the necessity of corresponding variables for establishing simulation validity. "A study conducted in a very different setting may be highly relevant if it has isolated a variable that is crucial in international relations" (p. 597). Chadwick (1966) used this approach by factor-analyzing core variables in the Inter-Nation Simulation and separately factor analyzing quantifiable real world indices assumed to correspond to the core variables in the simulation. Judgments about simulation validity may be made from a comparison of the factors and of the variables loading on each factor.

Sensitivity testing is a feature of variable-parameter validity. In repeated runs of a game or simulation the setting of a parameter or the range of values assigned a variable are systematically changed to determine what difference, if any, the alteration has on the operation of the model. Sensitivity analysis can be contrasted with internal validity. The latter involves repeated runs of an operating model in which all variables and parameters are given identical values in each replication. On the other hand, sensitivity testing involves the systematic altering of a selected variable or parameter through successive runs while all other properties are held constant. Sensitivity analysis revealed that an inappropriate parameter had been included in the Simulmatics Corporation simulation of the 1960 Presidential election (see Pool, Abelson, and and Popkin, 1965). In repeated simulations of that election the researchers varied "voter turnout," a parameter of the proportion of various types of voters that would not cast a ballot because they experienced cross-pressures. The investigators discovered that any value other than zero given to this parameter reduced the correspondence between simulation results and the actual election.

The variable-parameter approach has the advantage of isolating individual components of the simulation. It is thus possible to determine what particular features may be reducing the representativeness of the operating model. The use of simulation variables and parameters introduces the problem of establishing operational definitions for their real world counterparts. (The problem is by no means unique to variable-parameter validity criteria.) This task can be particularly troublesome when the definitions must correspond to a simulation variable or parameter that is either an analogue or a prototype intended to combine numerous features of the reference system. Two further difficulties are encountered with variable-parameter validity. First, an operating model consists of more than variables and parameters. It also involves the relationships between these properties. This relational aspect of a game or simulation receives only indirect exploration through variable and parameter validity. A second, practical restriction applies to sensitivity analysis. The procedure is quite laborious and for a complex model can be almost endless.

Event validity

A different validity approach employs "natural" events as criteria against which to compare outcomes occurring in the simulation. In the present context, "event" is defined to include patterns of behavior—such as volume of communication—as well as isolated occurrences. One exploratory exercise with a simulation of international politics attempted to replicate features of the crisis preceding the First World War (Hermann and Hermann, in press). The starting values of the simulation were set to represent the principal nations engaged in events leading to the war. Participants were selected and assigned roles on the basis of their similarity to historical figures on several personality traits judged to be prominent in the major decision makers in 1914. Events that resulted from the operation of the simulation were compared to those that transpired during the critical days in 1914 to determine

the degree of correspondence.
tensive series of runs, currently
ducted with the Inter-Nation
endeavors to structure the initia
parallel to the current world c
(Meier, 1963). The simulation is
several "years" into the future
sequence of events that will be
those actually emerging in the
the next few years. Similarly, w
outcomes projected by a sim
compared with the actual camp
event validity is involved (Po
and Popkin, 1965).

To the extent that events can
with the consequences or end p
simulation, they provide the ma
mediate relevance to the inves
interests in prediction. Becaus
most circumscribed event is t
result of interaction among nu
ments in the model, event valida
quite useful for checking the t
tion, that is, the composite set
tionships. By the same reasonin
event validation may be less us
covering the exact parts of a
model responsible for incongrui
the simulation events and those i
to which it is compared. One c
tion for the event approach to va
level of generality at which even
compared. In the simulation of
I, we can inquire about very m
Did war occur? At a much more
we might ask if the participant
ing Russia held a conference co
to the Russian Crown Council
and, if they did, what decisions
The large number of contribu
required for the occurrence of a
the outbreak of a world war de
cation of many features of realit
of which might reverse the final
macro event if it failed. This lin
ing suggests that the more gene
an event in the observable u
greater will be an operating mode
in reproducing that event. Or
hand, the proposition has bee
(for example, see Kaplan, Sk
Girshick, 1950; Sprout and Sp
that the more specific and minu

the degree of correspondence. A more extensive series of runs, currently being conducted with the Inter-Nation Simulation, endeavors to structure the initial conditions parallel to the current world configuration (Meier, 1963). The simulation is then played several "years" into the future to create a sequence of events that will be compared to those actually emerging in the world over the next few years. Similarly, when election outcomes projected by a simulation are compared with the actual campaign results, event validity is involved (Pool, Abelson, and Popkin, 1965).

To the extent that events can be equated with the consequences or end products of a simulation, they provide the material of immediate relevance to the investigator with interests in prediction. Because even the most circumscribed event is the probable result of interaction among numerous elements in the model, event validation may be quite useful for checking the total simulation, that is, the composite set of interrelationships. By the same reasoning, however, event validation may be less useful for discovering the exact parts of an operating model responsible for incongruities between the simulation events and those in the reality to which it is compared. One critical question for the event approach to validity is the level of generality at which events should be compared. In the simulation of World War I, we can inquire about very macro events. Did war occur? At a much more micro level we might ask if the participants representing Russia held a conference corresponding to the Russian Crown Council on July 25, and, if they did, what decisions were made. The large number of contributing factors required for the occurrence of an event like the outbreak of a world war demands replication of many features of reality—any one of which might reverse the final outcome or macro event if it failed. This line of reasoning suggests that the more general or macro an event in the observable universe, the greater will be an operating model's difficulty in reproducing that event. On the other hand, the proposition has been advanced (for example, see Kaplan, Skogstad, and Girshick, 1950; Sprout and Sprout, 1961) that the more specific and minute the event predicted, the less the probability of a successful prediction.

A second problem encountered with event validity is the selection of the appropriate dimension for comparing occurrences in a simulation and its reference system. For example, a particular event may be found in both the game and its reference system, but the temporal sequence may differ. Germany mobilized after Russia in 1914, but suppose a simulated Germany mobilized first in an attempted replication. Or consider a simulation event that is clearly different in substance from any counterpart in the observable universe, but appears similar in the effect it achieves. Finally, the problem arises of weighting events of different frequencies of occurrence. Is it as important that a simulation replicate a diplomatic message (a rather frequent class of events) as it is that it replicate the elimination of a head of state (a less frequent class of events)? These questions illustrate some of the issues that must be faced in establishing the similarity or dissimilarity of events in laboratory and natural settings.[4]

Hypothesis validity

In this approach, hypothesized relationships become the validity criteria. If X is observed to bear a given relationship to Y in the observable universe, then X' should bear a corresponding relationship to Y' in a valid operating model. Hypothesis validity differs from parameter-variable or event validity in that the criteria are not individual entities, but connections between two or more units. At least two kinds of relationships can be used in hypothesis validity. First, we can explore programmed relationships which are an integral part of the operating model. These relationships can be stated as researchable hypotheses; or occasionally, empirically verifiable propositions can be derived from them. A second kind of hypothesis applicable in this validity approach is independent of the programmed relationships contained in the operating model itself. Games or simulations can provide appropriate settings for investigating

[4] Further discussion of problems experienced with event validity can be found in Nardin (1965) and Targ and Nardin (1965).

hypotheses even though the relationship between the variables is not required for the operation of the model.

Work undertaken by Zinnes (1966) illustrates the approach to simulation validity using hypotheses independent of the model's programmed structure. In her analysis, 13 hypotheses were statistically tested using simulation and historical data. The hypotheses involved the relationships between such variables as perception of hostility, expression of hostility, frequency of interaction, and bloc or alliance membership. The historical data were obtained from a content analysis of diplomatic communications exchanged between the major European powers in the 1914 crisis. A similar content analysis was performed on the communications between simulated nations in an operating model which made no attempt to replicate the events surrounding the outbreak of World War I. Correspondence was found in 9 of the 13 relationships. Interpretations were advanced to explain the absence of comparability on the remaining four hypotheses. Although their immediate concerns are different, Rice and Smith (1964) also advocate the hypothesis approach to validity.

We should recognize that some hypotheses may be valid in a variety of human systems (families, unions, and so on). The fewer the systems in which a relationship might be expected to be verified (other than the one which the model represents), the more critical a test the hypothesis provides of the operating model's correspondence to a given reference system. This argument parallels a validity discussion of psychological tests. In that literature the need for discriminant as well as convergent validity has been recognized (Campbell and Fiske, 1959). That is, a test designed to measure one psychological concept (such as risk taking) should not correlate with another instrument intended to measure a totally different concept (such as intelligence). If this interpretation were extended to games or simulations, an operating model would be increasingly valid as its operation was distinguishable from systems which it was not intended to represent, as well as by evidence of its convergence with the performance of the intended reference system.

In the hypothesis approach to simulation or game validity we need to select a series of relationships engaging a variety of different simulation components. Diversity should be considered not only from the perspective of different substantive variables and relationships, but also in terms of the methods used for representing the reference system in the model. As we commented in the introduction, features of the observable universe can be represented in an operating model by reduction, transformation, and substitution. Observations should be made to determine if systematic differences in correspondence occur with these different means of representation.

Similarities in proposed approaches

The five approaches for investigating the validity of simulations and games share several characteristics. They all involve the comparison of some aspect of the operating model with some criteria or standards. In internal validity, the criteria are replications of the game or simulation; the model becomes its own standard for establishing stability and consistency. The criteria used in face validity are vague, but they exist as subjective impressions of the relevant aspects of "reality." Each of the remaining three approaches selects an aspect of the simulation or game and attempts to locate corresponding criteria in the reference system.

With the possible exception of face validity, these approaches all imply statistical analysis. Unlike the usual research enterprise in behavioral science, the validation of a model is improved when the null hypothesis cannot be rejected. In other words, a correspondence is suggested when no significant difference can be identified between the criterion and the simulation or game. The usual formulation of the hypothesis of no difference, however, does not provide much information on the relationship between the model and the criterion. Grant (1962) has proposed a reorientation of the null hypothesis so that failure to reject it constitutes an exact expression of the covari-

ation between data from the moc
observable universe.

Statistical analysis emphasizes
for measuring the validity criter
that an estimate of the fit betw
and criteria can be made. In addi
problem of measurement is the br
tion of what the criteria represen
way to pose this issue is to ask: A
tion validity criteria "reality?"
engaging in an extensive philoso
logue, we accept Churchman's
servation that validity criteria
"assertions" about what is assum
Just as the games or simulations a
mations of "reality," so are es
variables, parameters, or relations
from the observable universe b
analysis, interviews, factor analy
other method. This gives rise to
bility of the "hypothesis actually b
and yet being confirmed by evide
natural world and in the ex
laboratory" (Rice and Smith,
though all possible criteria against
can evaluate a simulation or gam
approximations, our confidence in
may increase if a corresponden
independent estimate of "reality"
lished.

If a validity criterion is subje
tortion—particularly when it mus
ured with our existing techniques
cannot consider a simulation valid
it is comparable to a single criter
lation and game validity must al
matter of degree. We may be more
of the validity of one simula
another if the former has been su
(and supported by) a greater
validity approaches and criteria. I
still remain a matter of degree. C
our discussion is similar to the ar
psychological test and measurem
over "construct validity." Constru
can be defined as validity deter
numerous, partial criteria. It "m
vestigated whenever no criterio
verse of content is acceptable as en
quate to define the quality to be r
(Cronbach and Meehl, 1955, p.
cause no one criterion encomp

rgence with the performance of the in-
nded reference system.

In the hypothesis approach to simulation
game validity we need to select a series of
ationships engaging a variety of different
nulation components. Diversity should be
nsidered not only from the perspective of
ferent substantive variables and relation-
ps, but also in terms of the methods used
representing the reference system in the
odel. As we commented in the introduc-
n, features of the observable universe
n be represented in an operating model by
duction, transformation, and substitution.
servations should be made to determine
systematic differences in correspondence
ur with these different means of represen-
ion.

nilarities in proposed approaches

The five approaches for investigating the
lidity of simulations and games share
veral characteristics. They all involve the
nparison of some aspect of the operating
odel with some criteria or standards. In
ernal validity, the criteria are replications
the game or simulation; the model be-
nes its own standard for establishing sta-
ity and consistency. The criteria used in
e validity are vague, but they exist as
ojective impressions of the relevant as-
ts of "reality." Each of the remaining
ee approaches selects an aspect of the
nulation or game and attempts to locate
responding criteria in the reference sys-
n.

With the possible exception of face
idity, these approaches all imply statisti-
analysis. Unlike the usual research enter-
se in behavioral science, the validation of
nodel is improved when the null hypothe-
cannot be rejected. In other words, a
respondence is suggested when no sig-
cant difference can be identified between
criterion and the simulation or game.
e usual formulation of the hypothesis of
difference, however, does not provide
ch information on the relationship be-
en the model and the criterion. Grant
62) has proposed a reorientation of the
l hypothesis so that failure to reject it
nstitutes an exact expression of the covari-

ation between data from the model and the observable universe.

Statistical analysis emphasizes the need for measuring the validity criteria in order that an estimate of the fit between model and criteria can be made. In addition to the problem of measurement is the broader question of what the criteria represent. Another way to pose this issue is to ask: Are simulation validity criteria "reality?" Without engaging in an extensive philosophical dialogue, we accept Churchman's (1961) observation that validity criteria are only "assertions" about what is assumed to exist. Just as the games or simulations are approximations of "reality," so are estimates of variables, parameters, or relationships drawn from the observable universe by content analysis, interviews, factor analysis, or any other method. This gives rise to the possibility of the "hypothesis actually being false, and yet being confirmed by evidence in the natural world and in the experimental laboratory" (Rice and Smith, 1964). Although all possible criteria against which we can evaluate a simulation or game are only approximations, our confidence in the model may increase if a correspondence with an independent estimate of "reality" is established.

If a validity criterion is subject to distortion—particularly when it must be measured with our existing techniques—then we cannot consider a simulation validated when it is comparable to a single criterion. Simulation and game validity must always be a matter of degree. We may be more confident of the validity of one simulation than another if the former has been subjected to (and supported by) a greater variety of validity approaches and criteria. But it will still remain a matter of degree. Once again our discussion is similar to the argument in psychological test and measurement circles over "construct validity." Construct validity can be defined as validity determined by numerous, partial criteria. It "must be investigated whenever no criterion or universe of content is acceptable as entirely adequate to define the quality to be measured" (Cronbach and Meehl, 1955, p. 282). Because no one criterion encompasses the totality of the test (in our case, an operating model), multiple incomplete criteria are required.

We have arrived at the position, then, that multiple validity criteria are needed because of the error in measurement and because of the recognition that criteria can be only assertions about "reality." The necessity for diverse validity criteria also may be a function of the kind of simulation. The more abstract are the units of a model (in the sense of removal from empirical referents) and the more simplified its system of interaction, the more difficult becomes the task of identifying relevant validity criteria. In this sense it may be less difficult to establish a given degree of correspondence between the international political system and a complex simulation than between the international system and a Prisoner's Dilemma game. A similar observation is made by Pruitt (1964), although he notes that "where a simpler situation can be found that nevertheless bears similarity to international relations, orderly relationships will more easily be identified than in the Inter-Nation Simulation" (p. 10).

By this point the reader who has conducted simulation and game exercises may protest that the requirement for multiple checks against validity criteria defeats one of his reasons for using operating models. A researcher may turn to simulation experiments because of the extreme difficulty he encountered in collecting some types of systematic data in the natural setting. If the simulation researcher is required to make numerous validity checks on the variables or hypotheses in which he is interested, why not initially explore them directly in the observable universe? Several responses can be made. First, other reasons exist for working with operating models as suggested by the various purposes described in the first section. Moreover, a person may work with an operating model even when areas of the model have been subjected to little systematic validation. When some elements of a model have been supported by validity operations, we may *cautiously* infer a degree of validity to those related elements which have not. This inference appears more justified the more similar or related the relatively unchecked components are to those for which some degree of verification exists.

As a final observation about the five validity approaches under examination, we suggest that their individual strengths and weaknesses may indicate a possible sequence in which they can be employed. In the initial construction of the simulation or game as well as during its operation, face validity is appropriate because of its ease and simplicity. To establish the degree of control and stability available in the operating model, internal validity should be conducted soon after the model's development.[5] With the knowledge of the degree of internal validity, we are prepared to investigate in a more detailed fashion the extent to which we can extrapolate from the simulation to its intended reference system. At this point, attention can be directed to outputs produced by the model. The outputs may be either events or research hypotheses. Unlike earlier validity steps, this phase lends itself to the simultaneous performance of ongoing research. If a game is designed to project election outcomes, then runs for this purpose can be conducted. If a simulation is to provide a setting for exploring judicial behavior, then hypotheses can be investigated. For purposes of validity some of the investigated relationships should be capable of confirmation or refutation by methods independent of the operating model.

In all probability some distributions of events or some kinds of hypotheses will produce results with unacceptable divergence between the operating model and the observable universe. Although these incongruous outputs may not pinpoint the inadequacy in the model, they should provide a diagnosis of the general area which seems unrepresentative. At this stage the validity inquiry turns to the inputs and intervening processes in the model—variable-parameter validity and programmed hypothesis validity. Because "a complex model can predict real-world outcomes correctly and yet be wrong in many details" (Pool, et al., 1965, p. 64), an investigator may wish to pursue validity approaches which focus on the internal structure of the model at an earlier stage in the operation of the simulation. With a complex model, however, examining all possible parameter settings and investigating all the assumptions can be an extremely formidable task. By deferring this stage until areas of trouble have been isolated, we have put boundaries on the otherwise endless task of sensitivity testing.

VALIDITY OF GAMES WITH HUMAN PARTICIPANTS

The intent of this essay has been to set forth some of the multiple validity problems associated with the use of operating models. We have described the impact of both purpose and criteria on validity. So far, we have not dealt directly with a set of recurrent questions about one kind of operating model —those which involve human participants. How can college students behave like experienced political leaders? How can American participants represent the cultural values of other societies? How can players reflect the considerations involved in a national decision to go to war, when their wars cause no death or destruction? The issue raised by these inquiries affects model validity as directly as do those of purpose and criteria.

A distinction between games and simulations based upon the introduction of humans as an element of the model is relevant to the validity problem. Although no consensus exists on the critical differences between the concepts "game" and "simulation," human participation is frequently associated with gaming. Operating models which represent human behavior symbolically and without actual human players have been characterized as simulations (for example, Dawson, 1962; Thorelli and Graves, 1964). Elsewhere, Hermann (1965) has suggested that the

[5] Must internal validity or reliability precede external validity (that is, comparison of the model to external standards)? The position taken in this discussion is that the model user should have some indication of the extent to which identifiable properties in his model, as opposed to unknown extraneous factors, are accounting for the occurrences in the game or simulation. This requirement does not mean that further validity tests cannot continue until the model produces identical outcomes in every run (phenomena that may be unlikely in the reference system). It does mean that before going on, the user should have some indication of the *degree* to which between-run variance is unexplained by the known qualities of the operating model.

necessity for participants in operat
is a function of the degree to w
governing the relationships bet
ments of the model are program
game only a limited number of
erning the potential relationshi
model are explicated before it is
contrast, the relationships betwee
in a simulation are extensively pr
during the construction of the mo
the rules for the detailed opera
model are not fully specified, hum
pants and administrators becom
to make such judgments during
running of the model. Conversely
programmed are the relationsh
operating model, the more restricte
the participant's role.

Why humans are used

Because the introduction of hu
ticipants in a model complicate
problems, it may be useful to re
reasons why human players are
When applied to the study of
science, the representation of h
havior—either of individuals or
social entities—is essential for any
model. Our present stage of k
however, often makes the detailed
tion of many behavioral processe
dinarily difficult. For example,
of jurisprudence who develops a
simulation may find himself oblig
struct programs for relationships a
familiar psychological variables
finds necessary in order to repre
judicial decision making, but whi
of immediate concern to his pr
search interests. To circumvent p
this type, some designers have i
the entire human organism into th
rather than use symbolic repre
of selected processes such as sele
ception and retention of informati
making, and affective and cogn

[6] Several investigators (Guetzkow
1957; Zelditch & Hopkins, 1961) have
that a small number of participants c
not only for face-to-face group pro
also to represent features of much la
units. For example, in his public opi
Davison (1961) has single individuals
the positions of mass interest groups.

cy in the model, they should provide
osis of the general area which seems
sentative. At this stage the validity
turns to the inputs and intervening
es in the model—variable-parameter
and programmed hypothesis va-
Because "a complex model can pre-
al-world outcomes correctly and yet
ng in many details" (Pool, et al.,
. 64), an investigator may wish to
validity approaches which focus on
ernal structure of the model at an
stage in the operation of the simula-
Vith a complex model, however, ex-
g all possible parameter settings and
ating all the assumptions can be an
ly formidable task. By deferring this
ntil areas of trouble have been iso-
e have put boundaries on the other-
dless task of sensitivity testing.

LIDITY OF GAMES WITH HUMAN PARTICIPANTS

intent of this essay has been to set
me of the multiple validity problems
ed with the use of operating models.
e described the impact of both pur-
d criteria on validity. So far, we have
lt directly with a set of recurrent
as about one kind of operating model
which involve human participants.
an college students behave like ex-
d political leaders? How can Ameri-
articipants represent the cultural
of other societies? How can players
the considerations involved in a na-
decision to go to war, when their
ause no death or destruction? The
ised by these inquiries affects model
as directly as do those of purpose
eria.

tinction between games and simula-
ased upon the introduction of humans
ement of the model is relevant to the
problem. Although no consensus
n the critical differences between the
s "game" and "simulation," human
ation is frequently associated with
. Operating models which represent
behavior symbolically and without
human players have been character-
simulations (for example, Dawson,
horelli and Graves, 1964). Elsewhere,
an (1965) has suggested that the

necessity for participants in operating models is a function of the degree to which rules governing the relationships between elements of the model are programmed. In a game only a limited number of rules governing the potential relationships in the model are explicated before it is played. By contrast, the relationships between elements in a simulation are extensively programmed during the construction of the model. When the rules for the detailed operation of a model are not fully specified, human participants and administrators become essential to make such judgments during the actual running of the model. Conversely, the more programmed are the relationships in an operating model, the more restricted becomes the participant's role.

Why humans are used

Because the introduction of human participants in a model complicates validity problems, it may be useful to review some reasons why human players are included. When applied to the study of behavioral science, the representation of human behavior—either of individuals or of larger social entities—is essential for any operating model. Our present stage of knowledge, however, often makes the detailed specification of many behavioral processes extraordinarily difficult. For example, a student of jurisprudence who develops a computer simulation may find himself obliged to construct programs for relationships among unfamiliar psychological variables which he finds necessary in order to represent basic judicial decision making, but which are not of immediate concern to his primary research interests. To circumvent problems of this type, some designers have introduced the entire human organism into their models rather than use symbolic representations of selected processes such as selective perception and retention of information, choice making, and affective and cognitive procedures. Similarly, the use of groups of participants as a direct part of the game can facilitate the problem of representing larger social units.[6] The point to be made here is not that symbolic representation of human behavior is impossible; on the contrary some important work is being done in this area (Abelson and Carroll, 1965; Gullahorn and Gullahorn, 1963). Rather, the observation is that the introduction of human participants frequently saves considerable cost and effort in the construction of an operating model and may avoid involving the designer in areas which are inadequately understood. Presumably the builder of an operating model has made a conscious choice. He has elected to adopt a method that permits the introduction of complex human processes which otherwise might be beyond inclusion in his model. In selecting this method he is aware that it may confound the results of the structured aspects of his model in ways that can be extremely difficult to identify.

A second reason for using human players is derived from the purposes for which a model is constructed. If the game is intended to serve a pedagogical function, an opportunity for student participation is frequently required. Thus, a simulated aircraft designed for pilot training must permit the would-be pilots to interact with the model. Similarly, a game designed for management training must allow the trainees to become participants. There are other objectives of operating models that require human players. For example, a firm may conduct a business game with a management group in an effort to discover new marketing programs. A government agency may wish to invite personnel from various related departments to participate in a game in order to establish some experience in interagency communication and coordination on a given type of problem solving. In each of these situations, an important function of the model is the involvement of individuals through direct participation.

In addition to the facility for representing human processes and the fulfillment of instructional purposes, human participation introduces a class of "real properties" into an operating model. Real variables in a model are identical to those which operate

[6] Several investigators (Guetzkow & Bowes, 1957; Zelditch & Hopkins, 1961) have contended that a small number of participants can be used not only for face-to-face group processes, but also to represent features of much larger social units. For example, in his public opinion game, Davison (1961) has single individuals represent the positions of mass interest groups.

in the system which the model is intended to represent. They can be contrasted with all other model variables which by transformation or substitution are symbolic representations of their designated counterparts in the reference system. The presence of real properties makes possible the operation of some relationships as they function outside the model. Moreover, as a class of real variables, human participation in gaming alters one aspect of the model validity problem. For those properties directly represented in the simulation as they exist in the reference system, validity can be more confidently assumed. As a minimum, human participants frequently increase the face validity—that is, the impression that aspects of reality are being reproduced. Kennedy (1962) has described games with human participants as individuals or group operatings in synthetic environments. As this description suggests, gaming would appear to shift the validity problem to the accurate representation of the environment in which the real properties—human participants—are to be embedded.[7]

Special validity problems

Although the third reason for using human participants in models is intended to reduce some validity problems, one must still be certain that the game players are representative of those in the reference system. "Representativeness" is one of the four special difficulties resulting from the use of human participants that we shall examine in the remainder of this section. In statistical terms the question is whether the game players and the reference system actors are drawn from the same population with regard to characteristics believed to be relevant to the model (such as intelligence, political experience, cultural values, and so forth). The problem corresponds to the issue in the psychological test and measurement literature called content validity—the adequacy with which a test samples the universe of attitudes, and so on, that it is designed to measure. This would appear to be the validity problem raised when one asks if high school students can act like national political leaders? To answer this question we must first establish the dimensions along which we require correspondence. Once the salient dimensions have been isolated, the degree of representativeness of selected participants becomes an empirical question capable of solution. Thus, the representativeness problem introduced by human participation in gaming seems manageable. Moreover, it should be recognized that this aspect of the validity problem exists even if all the properties in the simulation are symbolic.

A closely related validity problem created by the use of human participants is the assignment of players to unfamiliar roles. An illustration of this difficulty is represented in a game in which American civilians are instructed to "act like" Soviet military policy makers. When an individual assumes a role for which he has little or no information, the resulting behavior may involve major and systematic distortions from the behavior associated with that role in the reference system. Once this potential problem in human gaming is recognized, several methods can be employed to cope with it. The most obvious solution is to assign positions in a game only to those who are quite familiar with the role they are expected to play. Thus, in the MIT political-military exercise (Bloomfield and Whaley, 1965), the individuals who act as Soviet policy makers are usually government or academic specialists on Russian affairs. Even under these conditions it may be highly desirable for the player to explicate his basic assumptions about the behavior pattern to be associated with the role. When individuals with experience in a given position required by a game are not available, an alternative solution is possible. The game may be so structured that certain constraints and responsibilities

[7] Using humans as real properties may create disparities in time between the players and their environment. In many games time is greatly compressed (or occasionally expanded) so that, for example, a year is "played" in one hour. Although many dynamics of the game are accelerated, the real humans may be unable to escape the bounds of biological time which limit the speed with which they can read messages, communicate, and so forth. For a discussion of the requirements necessary for creating a synthetic environment for human participants, see Drabek and Haas (in press).

are imposed by the format of th
other words, the participant is in
act as himself. He is not encourag
his role on the basis of his conjec
some other actors would behav
role attributes are indicated by s
quirements built into the mode
stract positions found in the In
Simulation (Guetzkow, Alger, Br
and Snyder, 1963) illustrate this
Instead of instructing a partici
sume the dictatorial powers of
Stalin, aspects of the authoritari
relationship to his polity are bu
requirements of his office. Tech
role structuring include the use
for pursuing specified objectives
of the available communication ch
kind of information that can be de
received through each channel, an
Some subtle role characteristic
difficult to achieve in this way, but
tions must be weighed against t
expertise of the probable particip

An additional problem associat
use of human participants in gam
reliability, or what we have call
validity. The players and adminis
game, consciously or unconsc
likely to alter the rules governing
tion of the model in different tr
of the same game. Several co
emerge. First, as the freedom of
pants increases to introduce new
unrecognized, elements into the
task of establishing what vari
tributed to a given result bec
complex. Second, considerable b
variance is probable; that is, it b
tremely difficult to hold consta
ments and their interrelationsh
poses of replication. The result is
—that is, models in which the p
determine the rules of relatio
likely to have low internal valid
components of a model and the
ment are subject to uncontrolle
tion, then the task of estab
model's degree of corresponden
external reference system becom
meaningless. A correspondence to
obtained during one performanc

are imposed by the format of the model. In other words, the participant is instructed to act as himself. He is not encouraged to define his role on the basis of his conjecture of how some other actors would behave. Instead, role attributes are indicated by specified requirements built into the model. The abstract positions found in the Inter-Nation Simulation (Guetzkow, Alger, Brody, Noel, and Snyder, 1963) illustrate this alternative. Instead of instructing a participant to assume the dictatorial powers of a Hitler or Stalin, aspects of the authoritarian leader's relationship to his polity are built into the requirements of his office. Techniques for role structuring include the use of rewards for pursuing specified objectives, definition of the available communication channels, the kind of information that can be delivered and received through each channel, and so forth. Some subtle role characteristics may be difficult to achieve in this way, but the limitations must be weighed against the relevant expertise of the probable participants.

An additional problem associated with the use of human participants in games concerns reliability, or what we have called internal validity. The players and administrators of a game, consciously or unconsciously, are likely to alter the rules governing the operation of the model in different trials or runs of the same game. Several consequences emerge. First, as the freedom of the participants increases to introduce new, and often unrecognized, elements into the game, the task of establishing what variables contributed to a given result becomes quite complex. Second, considerable between-run variance is probable; that is, it becomes extremely difficult to hold constant the elements and their interrelationships for purposes of replication. The result is that games —that is, models in which the participants determine the rules of relationship—are likely to have low internal validity. If the components of a model and their arrangement are subject to uncontrolled modification, then the task of establishing the model's degree of correspondence to some external reference system becomes almost meaningless. A correspondence to a referent obtained during one performance of such a model may not recur in any future operations of the same model.

The last implication of using human participants which we shall consider introduces one of the most perplexing validity problems. Can players obtain motivation in a temporary, synthetic environment similar to that which they experience in the actual reference system? The issue is dramatized within a model by the representation of situations which in actuality would have severe consequences for the individual—that is, extreme physical or psychological deprivation. What does it mean to be financially bankrupt in a game or to engage in nuclear war? If the symbolic representation of such conditions in a game is known to the participants not to carry lasting significant effects, then it may be doubtful that players will manifest behaviors equivalent to that found in the reference system. Many commentators may be inclined to share Davis's (1966) conclusion that "the emotional overtones which always characterize international crises can seldom—if ever—be simulated in the laboratory" (p. 240).

Inducing an equivalent motive state may be a validity criteria that cannot be closely approximated for some kinds of circumstances in a laboratory setting. But several possible approaches to the problem can be described. The first alternative is to avoid using humans in operating models designed to explore situations which in the reference system involve extreme consequences for individuals. Thus, TEMPER (Apt, 1964), a computer simulation designed to explore various kinds of global conflicts, does not include human participants. This solution assumes that it is more likely for participants to display similar behavior when the actual situation to be represented does not invoke extraordinary rewards or punishments for the humans involved. A second approach is based on the selection of participants who have considerable ability to become totally involved in any environment into which they are introduced. The assumption in this alternative is that there are significant individual differences in the ability to become engaged in synthetic environments. Moreover, those persons who display this role-playing capacity are assumed to manifest the same behaviors as they would if the environment were not a temporary, synthetic one. Finally, situations demanding high involvement might be deferred until some predetermined level of involvement had been obtained. If participants in various runs of the model never reached the prescribed threshold, then the particular game might have to be discarded—at least for that purpose. Several indicators of involvement can be suggested. For example, several studies report that under moderate stress various kinds of performance increase, but as the intensity of the stress continues to rise performance falls off sharply. In an attempt to produce stress situations, the threshold might be an indication that performance had passed the peak and turned downward. Measures of ethnocentrism are also available. If a high degree of nationalism were identified in the reference system behavior, some minimal degree of ethnocentrism might be required before accepting data produced by a game. The particular criteria will depend on the kind of situation we attempt to replicate. One of the assumptions of this approach is that if a game can produce some critical thresholds of participant involvement, then the players' behavior may approximate what would occur under much more extreme conditions.

CONCLUSION

The primary issues discussed in this article may be summarized in four observations. (1) The validation of a simulation or game is always a matter of degree. Moreover, a given operating model may be relatively more valid by some criteria than by others. (2) The validation of an operating model cannot be separated from the purpose for which it is designed and conducted. Therefore, a simulation or game relatively valid for one objective may not be equally valid for another. (3) Given multiple validity strategies, several of the broadly applicable criteria may be reasonably applied in a particular sequence. Face validity can be used in the construction stages; internal validity can follow in the early trial runs; when research is being performed event and research hypothesis validity can be employed; if the previous validity checks indicate areas of poor correspondence, variable-parameter and programmed hypothesis validity can then be engaged. (4) The use of human participants in games significantly alters the required validation procedures. Although some major problems are reduced by this introduction of real properties, the net result would appear to make the estimation of validity more complex.

An examination of the game and simulation literature from which the references in this essay were drawn leads to a final conclusion. Validation questions, other than that of face validity, have yet to be explored for most operating models. Because a comprehensive investigation entails a variety of approaches, we are confronted with operating models whose degree of validity is largely unknown. Thus, insofar as their validation is concerned, it is premature to reject or accept the value of most simulations and games in the behavioral sciences.

REFERENCES

Abelson, R. P., & Carroll, J. D. Computer simulation of individual belief systems. *Amer. behav. Scientist*, 1965, 8, 24–30.

Apt, C. C. War gaming. *Int. Sci. Technol.*, 1964, 32, 29–37.

Barringer, R. E., & Whaley, B. The MIT political-military gaming experience. *Orbis*, 1965, 9, 437–458.

Bloomfield, L. P., & Whaley, B. The political-military exercise: A progress report. *Orbis*, 1965, 8, 854–870.

Boguslaw, R., Davis, R. H., & Glick, E. B. National policy formation in a less armed world. *Behav. Sci.*, 1966, 11, 43–61.

Brody, R. A. Some systemic effects of the spread of nuclear weapons technology: A study through simulation of a multi-nuclear future. *J. conflict Resolut.*, 1963, 7, 665–753.

Campbell, D. T. Factors relevant to the validity of experiments in social settings. *Psychol. Bull.*, 1957, 54, 297–312.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81–105.

Caspary, W. R. The causes of war in Inter-Nation Simulation-8. Unpublished manuscript, Political Science Dept., Northwestern University, 1962.

Chadwick, R. W. Developments in a partial theory of international behavor: A test and extension of inter-nation simulation theory. Unpublished doctoral dissertation, Political Science Department, Northwestern University, 1966.

Chapman, R. L., Kennedy, J. L., N
Bill, W. C. The Systems Resea
tory's air-defense experimer
Guetzkow (Ed.) *Simulation in s*
Readings. Englewood Cliffs, N.
Hall, 1962, Pp. 172–188.

Churchman, C. W. An analysis of th
simulation. Unpublished manu
agement Science Research Cen
sity of California at Berkeley,

Cohen, K. J., Cyert, R. M., Dill, W.
A. A., Miller, M. H., Van Wo
& Winters, P. R. The Carnegi
agement Game. In H. Guet
Simulation in social science
Englewood Cliffs, N.J.: P
1962, Pp. 104–123.

Cohen, K. J., Dill, W. R., Kuehr
Winters, P. R. *The Carnegie T*
ment Game. Homewood, Ill.: I

Cronbach, L. J., & Meehl, P. E. Const
in psychological tests. *Psychol*
52, 281–302.

Crow, W. J., & Noel, R. C. *The valid*
tion results. Unpublished rep
theon Corporation, Western
Sciences Institute, LaJolla, Cal

Davis, R. M. The international influ
Amer. Psychol., 1966, 21, 236–2

Davison, W. P. A public opinion gam
Quart., 1961, 25, 210–220.

Dawson, R. E. Simulation in the socia
H. Guetzkow (Ed.) *Simulati*
science. Englewood Cliffs, N.
Hall, 1962, Pp. 1–15.

Drabek, T. E., & Haas, J. E. Realis
tory simulation: Myth or me
Forces, in press.

Giffin, S. F. *The crisis game*. New Y
day, 1965.

Grant, D. A. Testing the null hypot
strategy and tactics of investiga
ical models. *Psychol. Rev.*, 196

Guetzkow, H. A use of simulation in
inter-nation relations. *Behav.*
183–191.

Guetzkow, H., Alger, C. F., Brody,
R. C , & Snyder, R. C. *Simula*
national relations. Englewood
Prentice-Hall, 1963.

Guetzkow, H., & Bowes, Anne E.
ment of organizations in a
Mgmt. Sci., 1957, 4, 380–402.

Gullahorn, J., & Gullahorn, Jeanne E
model of elementary social b
hav. Sci., 1963, 8, 354–363.

Hermann, C. F., Games and simulati
cal processes. Article prepared
tional Encyclopedia of the So
October, 1965.

Hermann, C. F., & Hermann, Marga
tempt to simulate the outbre
War I. *Amer. pol. sci. Rev.*, in

previous validity checks indicate areas
r correspondence, variable-parameter
rogrammed hypothesis validity can
e engaged. (4) The use of human par-
ts in games significantly alters the re-
validation procedures. Although some
problems are reduced by this intro-
n of real properties, the net result
appear to make the estimation of
y more complex.
examination of the game and simula-
erature from which the references in
say were drawn leads to a final con-
. Validation questions, other than
face validity, have yet to be explored
st operating models. Because a com-
sive investigation entails a variety of
ches, we are confronted with operat-
dels whose degree of validity is largely
wn. Thus, insofar as their validation
erned, it is premature to reject or
the value of most simulations and
in the behavioral sciences.

REFERENCES

, R. P., & Carroll, J. D. Computer simulation of individual belief systems. *Amer. Behav. Scientist*, 1965, 8, 24–30.

C. War gaming. *Int. Sci. Technol.*, 1964, 2, 29–37.

er, R. E., & Whaley, B. The MIT political-military gaming experience. *Orbis*, 1965, 9, 37–458.

eld, L. P., & Whaley, B. The political-military exercise: A progress report. *Orbis*, 965, 8, 854–870.

w, R., Davis, R. H., & Glick, E. B. National policy formation in a less armed orld. *Behav. Sci.*, 1966, 11, 43–61.

R. A. Some systemic effects of the spread f nuclear weapons technology: A study hrough simulation of a multi-nuclear future. *J. conflict Resolut.*, 1963, 7, 665–753.

ll, D. T. Factors relevant to the validity f experiments in social settings. *Psychol. Bull.*, 1957, 54, 297–312.

ll, D. T., & Fiske, D. W. Convergent and iscriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 6, 81–105.

, W. R. The causes of war in Inter-Nation imulation-8. Unpublished manuscript, olitical Science Dept., Northwestern niversity, 1962.

ck, R. W. Developments in a partial heory of international behavor: A test and xtension of inter-nation simulation theory. npublished doctoral dissertation, Political cience Department, Northwestern University, 1966.

Chapman, R. L., Kennedy, J. L., Newell, A., & Bill, W. C. The Systems Research Laboratory's air-defense experiments. In H. Guetzkow (Ed.) *Simulation in social science: Readings*. Englewood Cliffs, N.J.: Prentice-Hall, 1962, Pp. 172–188.

Churchman, C. W. An analysis of the concept of simulation. Unpublished manuscript, Management Science Research Center, University of California at Berkeley, 1961.

Cohen, K. J., Cyert, R. M., Dill, W. R., Kuehn, A. A., Miller, M. H., Van Wormer, T. A., & Winters, P. R. The Carnegie Tech Management Game. In H. Guetzkow (Ed.) *Simulation in social science: Readings*. Englewood Cliffs, N.J.: Prentice-Hall, 1962, Pp. 104–123.

Cohen, K. J., Dill, W. R., Kuehn, A. A., & Winters, P. R. *The Carnegie Tech Management Game*. Homewood, Ill.: Irwin, 1964.

Cronbach, L. J., & Meehl, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 52, 281–302.

Crow, W. J., & Noel, R. C. *The valid use of simulation results*. Unpublished report for Raytheon Corporation, Western Behavioral Sciences Institute, LaJolla, Calif., 1965.

Davis, R. M. The international influence process. *Amer. Psychol.*, 1966, 21, 236–243.

Davison, W. P. A public opinion game. *Pub. opin. Quart.*, 1961, 25, 210–220.

Dawson, R. E. Simulation in the social sciences. In H. Guetzkow (Ed.) *Simulation in social science*. Englewood Cliffs, N.J.: Prentice-Hall, 1962, Pp. 1–15.

Drabek, T. E., & Haas, J. E. Realism in laboratory simulation: Myth or method? *Social Forces*, in press.

Giffin, S. F. *The crisis game*. New York: Doubleday, 1965.

Grant, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychol. Rev.*, 1962, 69, 54–61.

Guetzkow, H. A use of simulation in the study of inter-nation relations. *Behav. Sci.*, 1959, 4, 183–191.

Guetzkow, H., Alger, C. F., Brody, R. A., Noel, R. C , & Snyder, R. C. *Simulation in international relations*. Englewood Cliffs, N.J.: Prentice-Hall, 1963.

Guetzkow, H., & Bowes, Anne E. The development of organizations in a laboratory. *Mgmt. Sci.*, 1957, 4, 380–402.

Gullahorn, J., & Gullahorn, Jeanne E. A computer model of elementary social behavior. *Behav. Sci.*, 1963, 8, 354–363.

Hermann, C. F., Games and simulations of political processes. Article prepared for *International Encyclopedia of the Social Sciences*, October, 1965.

Hermann, C. F., & Hermann, Margaret G. An attempt to simulate the outbreak of World War I. *Amer. pol. sci. Rev.*, in press.

Kahn, H. *Thinking about the unthinkable*. New York: Horizon, 1962.

Kaplan, A., Skogstad, A. L., & Girshick, M. A. The prediction of social and technological events. *Pub. opin. Quart.*, 1950, 14, 93–110.

Kelman, H. C. (Ed.) *International behavior*. New York: Holt, Rinehart, & Winston, 1965.

Kennedy, J. L. The system approach: Organizational development. *Human Factors*, 1962, 4, 25–52.

Kress, P. F. On validating simulation: With special attention to simulation of international politics. Unpublished manuscript, Political Science Dept., Northwestern University, 1965.

Meier, Dorothy L. Progress report: Event simulation project. Unpublished manuscript, Political Science Department, Northwestern University, November, 1963.

Nardin, T. An inquiry into the validity of a simulation of international relations. Unpublished manuscript, Political Science Dept., Northwestern University, August, 1965.

Pool, I. S., Abelson, R. P., & Popkin, S. L. *Candidates, issues, and strategies*. (Rev. ed.) Cambridge, Mass.: M.I.T. Press, 1965.

Pool, I. S., & Kessler, A. The kaiser, the tsar, and the computer. *Amer. behav. Scientist*, 1965, 8, 31–38.

Pruitt, D. G. Some comments on the use of simulation in the study of international relations. Paper presented at the Symposium on Psychology and International Relations, Georgetown University, Washington, D.C., June, 1964.

Rice, D. B., & Smith, V. L. Nature, the experimental laboratory, and the credibility of hypotheses. *Behav. Sci.*, 1964, 9, 239–246.

Sprout, H., & Sprout, Margaret. Explanation and prediction in internation politics. In J. N. Rosenau (Ed.) *International politics and foreign policy*. New York: Free Press, 1961, Pp. 60–72.

Targ, H. R., & Nardin, T. The Inter-Nation Simulation as a predictor of contemporary events. Unpublished manuscript, Political Science Dept., Northwestern University, August, 1965.

Thompson, P. D. *Numerical weather analysis and prediction*. New York: Macmillan, 1961.

Thorelli, H. B., & Graves, R. L. *International operations simulation*. New York: Free Press, 1964.

Zelditch, M., & Hopkins, T. K. Laboratory experiments with organizations. In A. Etzioni (Ed.) *Complex organizations*. New York: Holt, Rinehart & Winston, 1961, Pp. 464–478.

Zinnes, Dina A. A comparison of hostile behavior of decision-makers in simulated and historical data. *World Pol.*, 1966, 18, 474–502.

(Manuscript received May 12, 1966)