TESTS OF MODEL BEHAVIOR THAT CAN DETECT STRUCTURAL FLAWS: DEMONSTRATION WITH SIMULATION EXPERIMENTS

Yaman Barlas
Department of Systems Analysis
Miami University
Oxford, OHIO 45056 - USA

I. INTRODUCTION

Validation of System Dynamics (SD) models involves two general types of tests: Tests of model *structure* and tests of model *behavior.* Yet, since SD models are causal models, the essence of model validity lies always in *structural* validity: "Right behavior for the right reasons". (The nature of SD model validity has been discussed in various SD literature. For example, Forrester 1961 (Chapter 13); Forrester 1968; Forrester and Senge 1980; Bell and Senge 1980; Richardson and Pugh 1981 (Chapter 5 and 6) and Barlas 1985). Standard behavior tests, which compare the model-generated behavior to the observed *reference* behavior are generally "weak" in SD context, since they can not separate spurious behavior accuracy ("right behavior for the wrong reasons") from true behavior validity. Such tests provide no *structural* information. Structure tests, on the other hand, are "strong" tests, since they evaluate the model structure *directly*. But their their major drawback is that they are *informal, qualitative*, hence difficult to communicate. (See Forrester and Senge 1980, and Richardson and Pugh 1981 (chapter 5) for some specific behavior and structure tests). Thus, it seems like a "third type of test" would be most appropriate for SD validation purposes: Quantitative/formal behavior tests that *can* provide some structural information ("structurally-oriented" behavior tests). Interestingly, such tests already exist in the SD validation "repertoire". (For example, "Extreme Condition" simulations, "Behavior Sensitivity" testing, in Forrester and Senge 1980 and Richardson and Pugh 1981). But these tests are usually listed along with all other behavior tests, which undermines the major difference that exists between the former and the latter.

In this paper, using simulation experiments, we demonstrate that the "structurally-oriented" behavior tests and other standard behavior tests are different in a fundamental sense. We also show how the structurally-oriented behavior tests can help diagnose/remove structural flaws. Thus, we suggest that such tests be identified and analyzed by System Dynamicists in more detail. It is hoped that the tests will be improved, standardized and implemented as part of all the major SD simulation software.

II. EXPERIMENTAL PROCEDURE

In order to make controlled experimentation possible, we first build a model that describes the dynamics of a hypothetical epidemic disease. We treat this model as a "synthetic real system", and then build three different models of it, deliberately incorporating different structural flaws in each one of them. Through simulation experiments, we demonstrate that i) models with major structural flaws can generate highly accurate behavior patterns which would pass the "weak" behavior tests, and ii) the "structurally-oriented" behavior tests can detect such spurious behavior accuracy and help uncover the "hidden" structural flaws. Three structurally-oriented behavior tests will be illustrated: A- Extreme Conditions test: Assign extreme values to selected parameters, and compare the model-generated behavior to the observed (or "anticipated") behavior of the real system, under the same extreme condition. B- The Behavior Sensitivity test: Determine those parameters to which the model behavior displays high sensitivity, and ask if the real system would exhibit a similar high sensitivity. C- Modified-Behavior Prediction test: It is sometimes possible to find data about the behavior of some "modified" version of the real system (which could be due to parameter modifications, or structural

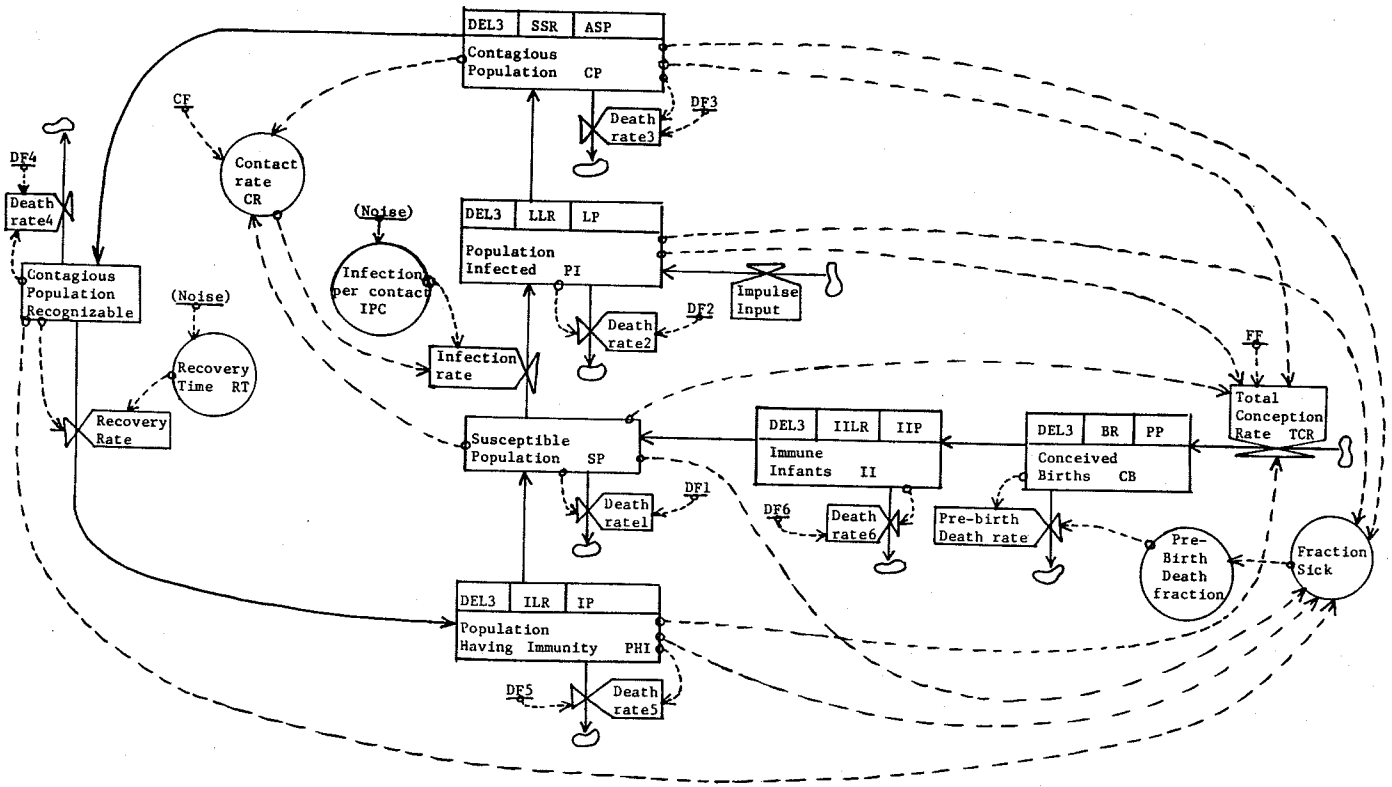

Figure 1. The Flow Diagram of the Synthetic Real System.

modifications). Then, test if the model is able to generate the same "modified behavior", when simulated with similar modifications. (All of the above three tests were first suggested by Forrester and Senge 1980. Our definitions and terminology given above are somewhat different from theirs, but these differences are merely conventional, not essential in any way).

III. THE SYNTHETIC REAL SYSTEM

As mentioned above, the synthetic real system used in the simulation experiments describes the dynamics of a hypothetical epidemic disease (Figure 1). The flow diagram consists of two major sub-structures: First, there is the loop that "recycles" the given population, through the various stages of the epidemic cycle: Susceptible Population --> Infected Population --> Contagious Population --> Contagious Population Recognizable --> Population Having Immunity --> Susceptible Population, which closes the loop. Infection is caused by the Contagious Population contacting ("Contact Rate") the Susceptible Population. This loop (on the left half of Figure 1) causes its variables to exhibit an oscillatory behavior pattern. As seen in Figure 2, the average period of the oscillations is about 32 months. (The randomness that is apparent in Figure 2 is obtained by using two Normal random variables in the model: Infections Per Contact and Recovery Time. The two behavior patterns depicted by the thin line and the thick one are generated by simulating the same system with two different noise seeds). The second major sub-structure of the model consists of the "population growth loops", depicted in the right half of Figure 1. Note that, except Contagious Population Recognizable, all population sub-groups affect the Total Conception Rate. These population growth loops create a positive trend in all population sub-groups, and thus the overall behavior of the model consists of epidemic oscillations superimposed on a trend (Figure 2). The output variable plotted in Figure 2 is "Contagious Population Recognizable" (CPR). To keep the experiments relatively simple, we will always assume that CPR is the only available output variable. (Due to space limitations, we skip the early transient portion of the behavior patterns, and provide a typical segment from t=80 to t=260). In addition to its reference behavior pattern, the synthetic system generates two modified patterns under two structural modifications: i) When all growth parameters are set to zero, the system exhibits oscillations around a *constant* mean, since all growth loops are in effect removed. The epidemic oscillations remain essentially unaffected. ii) When the Immune Period (IP) is set to an extremely large number -which practically means "permanent immunity"- then the regular epidemic fluctuations disappear. The modified behavior consists of "epidemic break outs", which occur with quite long

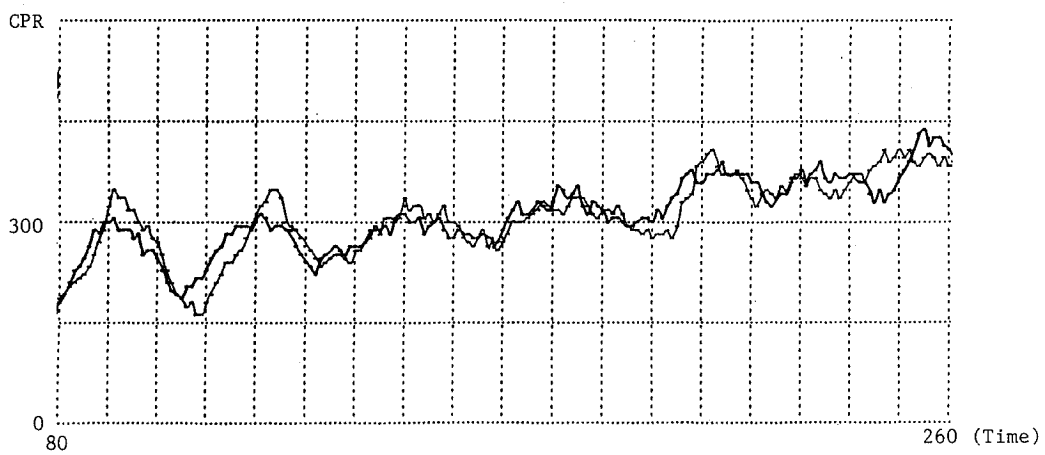

Figure 2. Reference Behavior Patterns of The Synthetic System, with Two Different Noise Seeds.

and variable intervals: The first epidemic starts at about t=10 and ends at t=40, the second one starts at t=130 and ends at t=180, the third one starts at t=205 and ends at t=260 etc. (This behavior is illustrated in Figure 4b by the solid line.Due to page restrictions, we are not able to provide the computer-generated graphs. We do not have space to provide the model equations either. The interested reader is referred to Barlas 1985 or 1989, for more detailed information). Finally, sensitivity of the behavior of the synthetic system to its parameters was examined. The fundamental *patterns* of behavior prove to be rather insensitive to most of the system parameters. Changing the delay times of the "recycling loop" (such as Immune Period or Recovery Time) by as much as 100% does *not* cause the behavior of the system to shift from one mode to another. The same is true for the parameters of the growth loops (such as Fertility Fraction). Changing the above parameters affects only the specific numerical values of behavior characteristics (such as increased amplitude, or shortened period). The only two parameters to which the system exhibits significant sensitivity are the ones belonging to the non-linear Infection Rate formulation:

$$IR.KL=IPC*CR.K, \text{ where } CR.K=CF*SP.K*CP.K \quad \ldots\ldots\ldots(1)$$

Increasing the values of IPC (Infection per Contact) or CF (Contact Fraction) by 50% or more results in huge oscillations that damp out immediately. Decreasing either of these by 50% results in unstable growing oscillations. Thus, the only parameters which the synthetic system is sensitive to are IPC and CF.

IV. SIMULATION EXPERIMENTS

In this section, we build three different models of the synthetic real system. Each model will have different structural errors/omissions, hence differing degrees of validity. We assume that the modeler observes the "real" behavior of the output variable CPR, for the period t=50 to t=200. (For simplicity, we assume that this is the only variable used in behavior testing. In an actual validation process, the tests discussed below would be applied to more than one variable).

Model-1. In this model, we assume that the analyst leaves out all the population growth loops (the entire right half of Figure 1). Lacking accurate information, the modeler perhaps attributes the observed growth of CPR to immigraton. Thus, the observed effects of the growth loops are replaced by a single external input function, estimated to be about 23 people/month. All other equations and parameter values are exactly the same as the real system's (except the noise sequences and the observation errors). The initial conditions at t=50 are also assumed to be known to the modeler. When Model-1 is simulated, CPR exhibits the behavior pattern represented by the thin line in Figure 3. (The thick line corresponds to the "real" behavior, with 3% observation erors). The agreement between the "real" and Model-generated patterns is obviously very good. Comparing Figure 3 to Figure 2, we see that even the real system itself does *not* replicate its own behavior much better than Model-1 does with a different noise seed! The behavior of model-1 will pass all standard ("weak") tests of behavior comparison. (Such as comparing the periods, amplitudes, trends etc. For detailed results of these, see Barlas 1985). Next, we subject Model-1 to the three structurally-oriented behavior tests:

A- Extreme Condition Test: As an extreme condition, we set RTAVG (average recovery time) to a very large number (like 1000000), which in reality means "no recovery". The resulting behavior of Model-1 is represented by the dashed line in Figure 4(a). This consists of a sharp increase followed by mild, long cycles superimposed on a trend. (The behavior patterns illustrated in Figures 4 through 7 are not precise, because they were hand-drawn in order to conserve space. Yet, they serve their purpose by conveying the fundamental *patterns* of behavior). The behavior of the synthetic real system, under the same extreme condition, is drastically different (solid line in Figure 4a): It is a sharp increase, followed by a steady decline (caused by deaths). In reality, in order to be able to apply such a test, the analyst would have to anticipate (using expert opinion, theory and data in established literature etc) the behavior of the real system under the extreme

condition. Once the test is applied, and the model fails it, the analyst must use reasoning to figure out the cause of failure: When RTAVG is made extremely large in the synthetic reality, CPR a constitutes a very large proportion of the overall population. This, in turn, results in drastically lower birth rates, because CPR is a non-reproductive group in the system. That is why CPR first shoots up, and then declines throughout the simulation. In Model-1, however, the net birth rate is independent of CPR, since the former is formulated simply as an external input.

B- Behavior Sensitivity Test: The sensitivity of Model-1 to its parameters is essentially the same as that of the synthetic real system. The model passes this test

C- Modified-Behavior Prediction Test: As mentioned above, the synthetic system generates the "recurrent epidemic" type of behavior (Solid line in Figure 4b), when the immunity is made permanent by setting IP (Immune Period) to a very large number. Assume that this "modified" behavior of the system was known to us. Apply the same modification to Model-1, which exhibits the behavior represented by the dashed line in Figure 4b. The only difference between the two graphs is that, in the "real" behavior the time interval between epidemic breakouts becomes shorter and shorter, whereas in the model it remains constant in the model

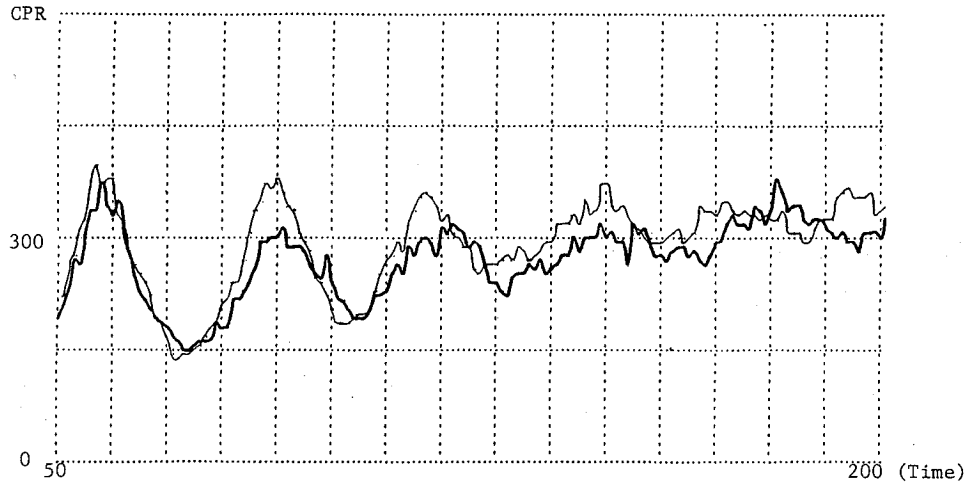

Figure 3. The Behavior Pattern Generated by Model-1 (Thin line), and the one Generated by the Synthetic Real System (Thick Line).

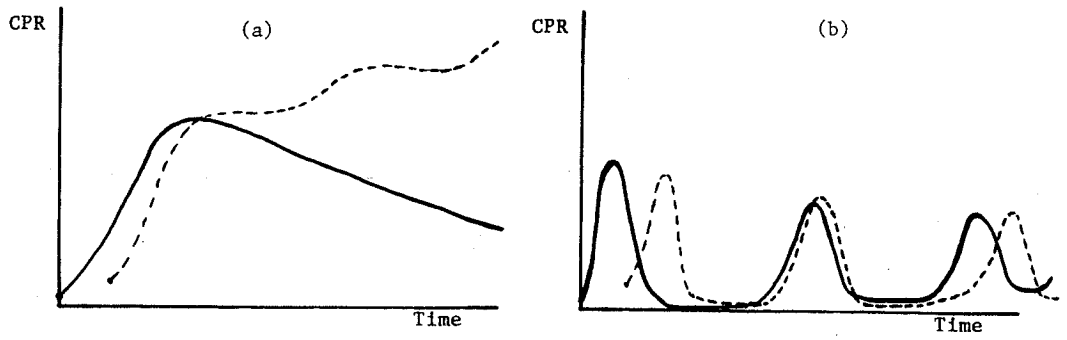

Figure 4 . Results of two "Strong" Behavior Tests for Model-1: (a) Extreme Condition Test with RTAVG=1000000 ("No Recovery"), (b) Modified-Behavior Prediction Test with IP=1000000 ("Permanent Immunity").
(Solid Line: "Synthetic Real" Behavior, Dashed line: Model-generated Behavior).

generated behavior. (Because the shortening of the interval is caused by the ever-increasing population growth rate, which is merely a constant in Model-1). Whether this subtle difference in the modified-behavior prediction would cast doubt on the usefulness of Model-1 would depend on its intended *purpose*.

Model-2. In this experiment, we incorporate a much more serious modeling error than the previous one: The immunity acquired after having the disease is assumed to be *permanent*. Thus, the crucial "recycling" loop (left half of Figure 1) is broken in Model-2. The analyst, who starts out with this incorrect aassumption of permanent immunity, later observes that the real behavior is oscillatory. To illustrate a bad modeling practice, let us assume that the analyst then tries to "fine tune" the remaining structure of Model-2, in order to generate oscillations out of it. One way of achieving this is to make the population growth rate high enough, which would create enough inflow into the Susceptible Population, yielding oscillations similar to the observed ones. Naturally, most of the parameters and initial conditions of this model would have to be "ad-hoc" fine-tuned in order to obtain the "right" oscillations (see Barlas 1985, for specific details). At the end Model-2 manages to generate a CPR behavior very similar to the one exhibited by the "observed" CPR in Figure 2. (Due to space restrictions, we are not presenting this particular graph). Once again, after enough "fine-tuning", the behavior of Model-2 passes the "weak" tests of behavior comparison, in spite of the fact that it has a very major structure error (see Barlas 1985). We next apply the three structurally-oriented behavior test to uncover the structural inadequacy of Model-2:

A- Extreme Condition Test: We repeat the extreme condition test done for Model-1: Set RTAVG to a very large number ("no recovery"). We already know that, under this extreme condition, the synthetic system generates a CPR behavior that consists of a sharp increase followed by a uniform decline (Solid line in Figure 5a). When the same extreme condition is applied to Model-2, CPR generates a behavior pattern very similar to the base run: Oscillations (with a period of about 35) superimposed on a positive trend (Figure 5a, dashed line). That is, the fundamental pattern of behavior of Model-2 changes very little under this extreme condition. The reason is obvious: Since Model-2 already assumes permanent immunity, making RTAVG extremely large does *not* break any loop! The population growth loops which are responsible for the oscillatory trend behavior of Model-2 are still active. (The growth of CPR becomes naturally much larger since there is no recovery under the extreme condition). Model-2 failing this extreme condition test would be taken by the modeler as an indication of a structural error. Investigating the *cause* of failure would further suggest the *location* of the structural error.

B- Behavior Sensitivity Test: As an example of sensitivity testing, we decrease the value of FF (Fertility Fraction) by 50%. As a result of this cahnge, the only noticeable change in the behavior of the "real" CPR is one of lowered trend slope (Figure 5b, solid line). That is, the behavior *pattern* of the synthetic system is insensitive to this particular parameter. But in Model-2, when the same parameter change is introduced, we see

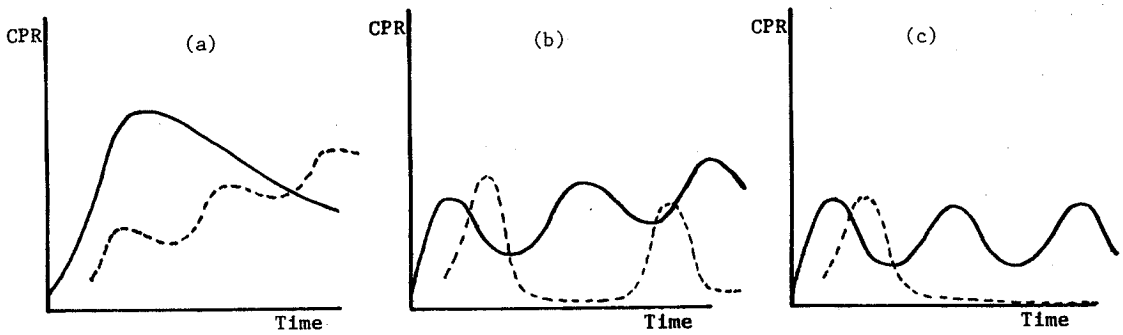

Figure 5. Results of:(a)Extreme Condition Testing with RTAVG=1000000, (b) Sensitivity . Testing with a 50% Decreasing of FF, (c)Modified-Behavior with "zero growth". (Solid Line: System Behavior, Dashed Line: Model-generated Behavior).

that the behavior of CPR changes drastically: Its period increases from 35 to 60, and it looks more like a "recurring epidemics" type of behavior (see Figure 5b, dashed lines). This extreme and improper sensitivity of Model-1 to FF would suggest that a the model may be lacking a major structure, and that the base behavior was obtained by excessive "fine tuning". More specifically, the test suggests that the excessively fine-tuned sub-structure (growth loops) includes the parameter FF.

C- Modified-Behavior Prediction Test: A modified-behavior of the synthetic system mentioned earlier is oscillations around a *constant* mean (Solid line in Figure 5c). The system exhibits this behavior when there is no population growth (i.e. the Fertility Fraction FF and all death fractions are 0). The Modified-Behavior Prediction test would ask if Model-2 is able to generate the same modified-behavior. When the "zero- growth" modifications are done in Model-2, the resulting behavior of CPR is a single growth-then-decline (dashed line in Figure 5c). The obvious failure of Model-2 comes from the fact that the population loops in this model are "fine-tuned" to generate the oscillations, and when these loops are removed, the oscillations completely disappear. This Modified-Behavior Prediction test would suggest to the analyst that Model-2 lacks the correct sub-structure that is responsible for the oscillations in the behavior of the real system.

Model-3. Finally, in a third model of the synthetic system, instead of incorporating a single obvious structural error, we incorporate many small "typical" modeling errors/omissions. Such errors would include neglecting some minor loops, replacing their anticipated effects by functional (TABLE) relationships, leaving out some of the accumulations or delays, representing some high-order delays with first-order delays etc.The following structural errors/omissions were incorporated in Model-3: i) Two third-order delays, Population Infected and Contagious Population are left out, and their effects are approximated by an increased Recovery Time and Immune Period. ii) The two successive delays, Immune Infants and Conceived Births are replaced by a single one, the length of which is equal to the sum of the two original delays. iii) All the minor loops representing the "death rates" are removed. Their overall effect is estimated and represented by a decreased

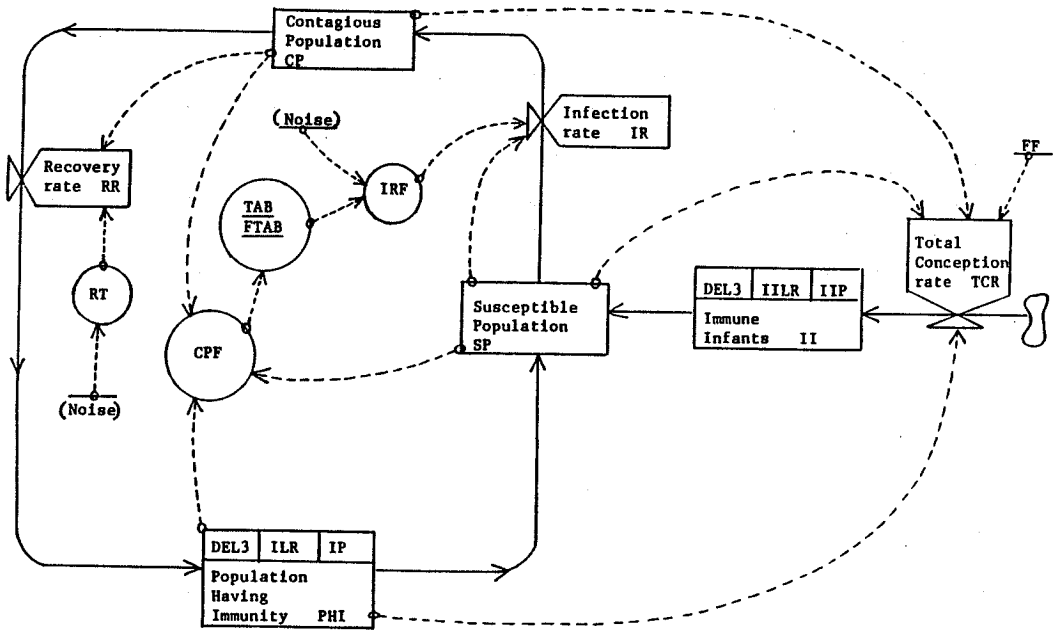

Figure 6. The Flow Diagram of Model-3.

average "net" growth rate. iv) The critical non-linear Infection Rate (see eq. (1) above) is changed. Such an alternative formulation is also needed to compensate for the effects of the two missing delays, Population Infected and Contagious Population. The Infection Rate in Model-3 is given by:

IR.KL=IRF.K*SP.K,

where the new variable Infection Rate Fraction (IRF) depends on Contagious Population Fraction(CPF):

IRF.K=TABHL(FTAB,CPF.K,0.0,0.16,0.02)+NOIS1.K

FTAB=0.0/0.020/0.040/0.065/0.090/0.110/0.125/0.140/0.150

CPF.K=CP.K/(CP.K+PHI.K+SP.K)

The complete Flow Diagram of Model-3 is given in Figure 6. Note that this model has considerable less detail and fewer variables than both the synthetic system, and the two previous models of it. In Model-3, we assume that the main output variable is CP, since CP and CPR are combined into a single variable. The important feature of Model-3 is that, although it is quite small and simple, it does contain (approximately) the two major sub-structures of the Synthetic System responsible for the oscillatory and trend patterns. It will therefore present subtler validity questions than the previous two models. When Model-3 is simulated under normal conditions, its behavior pattern is once again virtually indistinguishable from the behavior of the synthetic system (for complete list of equations and behavior patterns, see Barlas 1985). Thus, the base-run behavior of Model-3 easily passes the standard ("weak") behavior comparison tests (Barlas 1985). Next, we apply the three structurally-oriented behavior tests:

A- Extreme Conditions Test: As an extreme condition test, when RTAVG is set to 1000000 in Model-4, CP exhibits a persistent growth pattern (dashed line in Figure 7a). We know from previous experiments that, under the same extreme condition, the synthetic real system exhibits a growth-then-decline pattern (solid line in Figure 7a). The structural reason behind this failure of Model-4 is the fact that it does not distinguish between CP and CPR (Contagious Population Recognizable, which is non-reproductive in the synthetic system). In Model-4, both the Infection Rate and the Birth Rate depend primarily on CP and SP, and since an extreme increase in RTAVG raises the average level of CP, the resulting behavior is a persistent growth. This extreme condition test warns us that, depending on the intended use of the model, the CP/CPR distinction may prove to be a crucial one.

B- Behavior Sensitivity Test: As a sensitivity test, we double the value of RTAVG. Compare the solid and dashed curves of Figure 7b: The resulting behavior of the synthetic system is one of *stronger* oscillations with an increased slope (about 1.2). Model-4, on the other hand, exhibits very mild oscillations superimposed on a very strong slope (about 2.0). Model-4 displays a "wrong" sensitivity as a result of doubling RTAVG: Instead of exhibiting stronger oscillations, the latter almost disappear. This is again caused by the fact that no CP/CPR

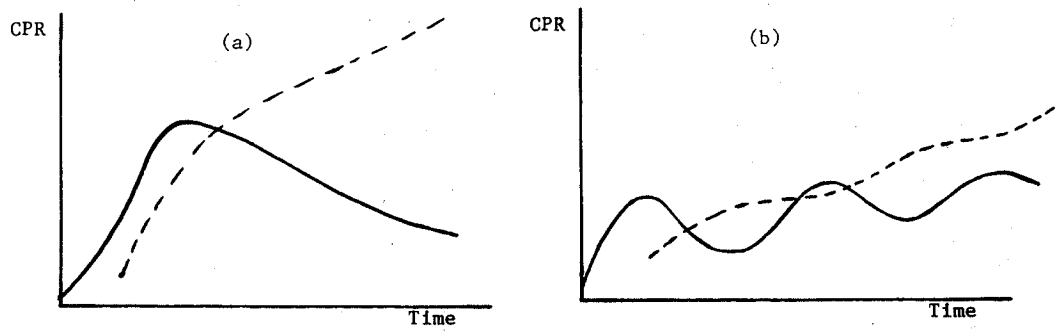

Figure 7. The Model-generated (Dashed line) and System-generated Behavior Patterns (solid line): (a) Under the Extreme Condition of RTAVG=1000000, (b) Sensitivities to Doubling the value of RTAVG.

distinction is made in Model-4. Since Infection Rate directly depends on CP, upon doubling RTAVG, the inner loop that generates the Infection Rate dominates the larger (recycling) loop which is the main machanism that creates the oscillations. As another sensitivity test, we increase the Immune Period (IP) by 50%. The synthetic system displays very little sensitivity to this. The resulting behavior is essentially the same as the base run, with slightly stronger oscillations, and slightly longer periods. Model-4, on the other hand, exhibits too much sensitivity, the resulting behavior consisting of very large and *growing* oscillations. This high and improper sensitivity to IP indicates that the order of Model-4 is too low and/or some important loops are missing. Indeed, Model-4 has too few accumulations and delays around the recycling loop.

C- Modified-Behavior Prediction Test: Model-4 is able to generate the the constant-mean oscillatory mode, by setting FF=0.0. With some minor problems, it is also able to generate the basic shape of the "recurrent" epidemic mode of behavior, by setting IP to a very large number.

Let us finally emphasize that, we do not suggest that models be simply "rejected" or "accepted" upon applying the above (or other) tests. To try to find a "yes or no" answer to model validity is not a constructive path. Instead, models should be viewed as "less" or "more" valid with respect to a given problem and purpose. In fact, one of the strongest arguments for the structurally-oriented behavior tests would be that they can help the analyst discover and remove the structural flaws behind a behavior inadequacy, and help increase the validity of a given model.

V- CONCLUSIONS

The above experimental results demonstrate that there are two fundamentally different types of behavior validation tests: "Weak" behavior tests compare the behavior of the model to a given reference behavior of the real system, and can not separate true behavior validity from spurious behavior accuracy. Such tests provide no structural information. Structurally-oriented "strong" behavior tests, on the other hand, can provide some structural information, hence help uncover the structural flaws of a given model. Experiments illustrate three such tests: Extreme Conditions test, Behavior Sensitivity test and Modified-Behavior Prediction test. Results show that these (and other similar) structurally-oriented behavior tests can be very useful in enhancing the degree of validity of a given model. We suggest that such tests be categorized and specific tests be analyzed by System Dynamicists in detail. It is hoped that structurally-oriented tests will be improved, formalized and eventually implemented as part of all the major SD simulation software.

REFERENCES

Barlas, Yaman. 1985. Validation of System Dynamics Models with a Sequential Procedure Involving Multiple Quantitative Methods (Ph.D. Dissertation), Georgia Institute of Technology, Atlanta, Georgia .

Barlas, Yaman. 1989. "Multiple Tests for Validation of System Dynamics Type of Simulation Models", European Journal of Operations Research, Forthcoming.

Bell, James A. and Senge, Peter, M. 1980. "Methods for Enhancing Refutability in System Dynamics Modeling", in Forrester Jay. W. et al., eds., System Dynamics, North-Holland, NewYork.

Forrester, Jay, W. 1961. Industrial Dynamics, MIT Press, Cambridge, MA.

Forrester, Jay, W. 1968. "A Response to Ansoff and Slevin", Management Science, Vol. 14, pp. 601-618.

Forrester, Jay, W. and Senge, Peter, M. 1980. "Tests for Building Confidence in System Dynamics Models", in Forrester, Jay. W. et al. eds., System Dynamics, North-Holland, New York.

Richardson, George P. and Alexander Pugh, III. 1981. Introduction to System Dynamics Modeling with DYNAMO, MIT Press, Cambridge, MA.