PARTIAL-MODEL TESTING AS A VALIDATION TOOL FOR SYSTEM DYNAMICS

Jack B. Homer
System Dynamics Group
Sloan School of Management
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This paper discusses an approach to model refinement which involves testing the behavior of individual pieces of a model in response to empirical input data for comparison with empirical output data. Partial-model tests should be used for selecting formulations or estimating parameters only when appropriate case-specific or logical information is not available for this purpose. The smaller the model components used for partial-model testing, the more likely it is that the model will prove useful for anticipating events outside the historical experience and the less likely it is that observed behavior will be incorrectly attributed to certain relationships or parameters. Thus, from the standpoint of structural validity, partial-model testing is an improvement over whole-model testing for the purpose of structural adjustment. The paper presents a detailed example of partial-model testing in the context of a generic model of the evolving use of a new medical technology. Specifically, the technique is used for adjusting and validating a model subsystem that can explain why the reporting of clinical information on cardiac pacemakers has been marked by regular oscillations over time.

1. Introduction

Validation of a system dynamics model can be viewed broadly as a process of demonstrating that both the structure and behavior of the model correspond to existing knowledge about the system under investigation.[1] [2] Matching empirical reference mode data is frequently the focus of behavioral validation efforts, although the production of unexpected yet believable patterns of behavior can also enhance the perceived usefulness of a model.[2][3] Reproduction of empirical reference modes is most con-

2

vincing when the model has been constructed and its parameters selected
on the basis of detailed information concerning pieces of the system which
are linked closely in time and space, rather than from correlations appar-
ent in aggregate time-series data. Ideally, all elements of structure
should be based on information at a level more detailed than that of the
model itself.[4] The more solid this foundation of data "below the level
of aggregation of model structure" is, the more likely it is that the model
will prove useful for anticipating events outside the historical experience
and generating accurate insights for policymakers. Conversely, to the ex-
tent formulations and parameter values are picked with the sole aim of
duplicating aggregate time-series data, the greater is the risk of producing
a structure which breaks down outside of the historical context and leads to
faulty conclusions.

Although one would like to formulate structures and estimate para-
meter values entirely on the basis of disaggregate case-specific data,
this is often impossible in actual practice, where time and resource con-
straints limit the effort that can be devoted to background research.
Models of "soft" systems, in which many of the quantities in question are
not directly measurable, pose particularly severe difficulties as far as
the collection of useful "micro" data is concerned. Of course, one can
often use logic or knowledge gained from general experience to fill in the
structural gaps left by empirical research; indeed, "educated guesses" are
part and parcel of the model-building process. But even after case-speci-
fic research and modeler's judgment are applied to construct a model, some
small number of gaps may remain. These gaps may be of two types: 1) Al-

3

ternative formulations for a given relationship may seem equally appealing,
and 2) one may be unable to establish a reasonably narrow range of accept-
able values for a given parameter. In such situations, one is thrown back
upon the use of aggregate time-series data as a guide to structure. This
use of time-series data means that a certain amount of curve-fitting will
be done in order to select formulations or parameter values; the purity of
a full a priori approach is thereby sacrificed. But this shortcoming need
not deal the validation effort a fatal blow; it is possible to test and
adjust a model without forcing or "fixing" the final result. Judicious
use of the technique of partial-model testing is the key.

2. The Technique: Description, Rationale, and Strategy

A partial-model test involves simulating the behavior of a func-
tional component of the model, which may be as small as a single equation,
in response to empirical input data for comparison with empirical output
data. The comparison with output data may be made by eye or with statis-
tical methods, as is true of any other data-based testing scheme.[5][6]
One begins with a guess as to which formulation or range of parameter
values is likely to provide an acceptable fit to the historical data,
where "goodness of fit" is defined relative to the model's purpose.[1]
If the initial fit is not acceptable, the uncertain structure of interest
is then adjusted until one finds a formulation or range of parameter values
for which an acceptable fit is obtained. (If the component being tested
contains n uncertain parameters, then one searches for a region in the rele-
vant n-dimensional parameter space for which the fit is acceptable.) Note
that formulations and parameter values will always be constrained to some

4

extent by real-life considerations; all pieces of structure must make sense even if they are uncertain. For instance, logical considerations will often dictate that a parameter value not be negative or that it be less than some maximum value.

Partial-model testing should be viewed as preferable to whole-model testing for the purpose of selecting or estimating pieces of structure. This follows from the idea that one would like to make such choices in a way that is consistent with available information but does not inappropriately "fix" the final simulation results. Indeed, as the size of the structure and the number of uncertain parameters in any single test increase, the potential for structural misspecification or misattribution of behavior to particular parameters also increases. For example, if two uncertain time constants together determine the frequency of an observed oscillation, estimation of both parameters in a whole-model context could lead to offsetting estimation biases.[4] Such misestimation could prove to be a critical flaw if the dominant real-life behavior has the potential of shifting from the original oscillation to some other mode in which one or both of the time constants still play a role. This problem could conceivably be avoided by estimating each uncertain parameter independently, in its own partial-model test. In general, the idea that uncertain formulations and parameter values should be established as independently as possible of the full model's structure suggests that the pieces of structure used in partial-model testing be as small as the available empirical data permit.

The component-by-component approach advocated here makes only partial use of all available data and structure in any given test. This means that

5

a simulation of the whole model using parameter values established via partial-model testing will do no better and will generally do worse in terms of fit than a simulation in which a "full information" approach to parameter estimation is taken.[7] However, this is precisely what one would expect from a technique that seeks to give greater certainty to structure without "fixing" the final simulation results. The smaller the functional component being tested and the smaller the number of parameters adjusted to improve that component's response to input data, the closer one comes to the ideal of determining structure on the basis of focused considerations of closely-related phenomena rather than the full range of observed macro-behavior.

The fact that partial-model testing does not guarantee the desired behavior of the full model implies that some problems in formulation or parameter estimation may not be revealed until the full model is simulated. Indeed, in actual practice it seems that some degree of adjustment and fine-tuning of the full model is inevitable and may even be the sign of a rich model—one in which the various components are coupled closely together. This suggests that partial-model testing might best be viewed as one step in an overall strategy of model development and testing which is intended to maximize both structural and behavioral validity. In the interest of structural validity, formulations and parameter values should be based on disaggregate case-specific data or other a priori considerations wherever possible, and partial-model testing used only as a secondary technique when the initial hypotheses fail to produce the desired result. Similarly, whole-model adjustments should be made only if partial-model tests have failed to reveal model inadequacies which become apparent when

the full set of feedback loops is simulated intact. Not only does such

a hierarchical approach to model-building maximize the likelihood that

the model will be useful for examining events outside the historical con-

text (since the structure is developed as independently as possible of the

full range of observed behavior), but it may also be the best general ap-

proach to troubleshooting during model development. This is simply because

it is easier to isolate the source of inappropriate behavior when it is seen

in a small piece of structure than when it is seen in a larger and more

complex structure. Thus, while "passing" a partial-model test does not

guarantee the appropriateness of a given piece of structure, "failing" the

test immediately identifies an area where improvement is needed and can

prevent much of the head-scratching involved in troubleshooting a large,

complex model.

3. Applying the Technique: An Example

3.1 Background

     The remainder of this paper is devoted to presenting an example of

partial-model testing in the context of a generic model of the evolving

use of a new medical technology.[8] This model was developed for the pur-

pose of generating insights useful to government policymakers regarding

the complex process by which new medical technologies are disseminated,

improved, and controlled. The model focuses on the actions of physicians,

who recommend, use, and evaluate the technology, and manufacturers, who are

responsible for promoting and modifying the technology. The full model con-

sists of approximately 150 equations, including over 30 levels and delays.

It was developed over a three-year period and its structure based on an

extensive review of the pertinent literature as well as in-depth case

studies of two different medical technologies.

     One of the cases studied was the implantable cardiac pacemaker, a

sophisticated electronic device used for restoring or maintaining a normal

heartbeat in certain patients who suffer from chronic heart rhythm problems

(arrhythmias). Ever since the first implantation in 1960, the field of

pacing has grown steadily, despite some early resistance by physicians to

the unusual new device. As Figure 1 indicates, the number of patients with

pacemakers ("recipients") in the U.S. increased from about two thousand in

1965 to over twenty thousand in 1970 to over one hundred thousand in 1975,

and by 1980 stood at just over 350 thousand. This growth is only partially

attributable to acceptance of the technique, which was essentially complete

by 1971. Rather, the rapid growth of the 1970s is almost fully a reflec-

tion of expanding clinical indications for pacing, which came in response

to several important technical developments which made the device both

safer and more broadly applicable.

     Figure 2 shows how the fraction of heart rhythm patients considered

eligible for pacing increased tenfold from less than two percent during the

early 1960s to nearly seventeen percent by 1980. Originally, only patients

with severe symptoms (dizziness, fainting, seizures) and at high risk of

sudden death were considered appropriate candidates. But as physicians

became bolder in their use of the device, they expanded the eligibility cri-

teria to include certain less symptomatic (sometimes asymptomatic) patients

as well as a whole class of patients with symptoms but with little risk of
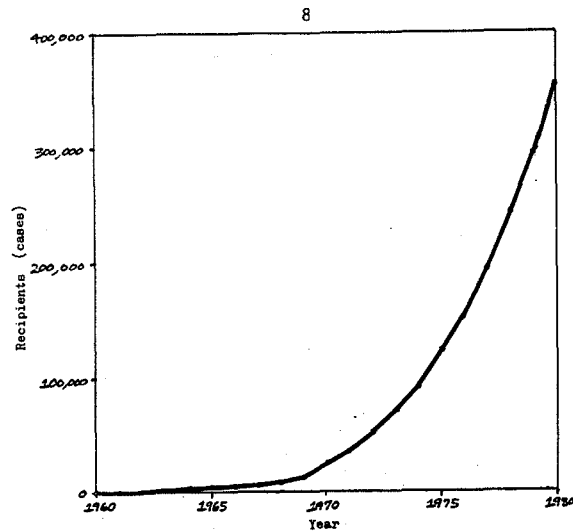
sudden death.[9]

8

Figure 1. Patients with Implanted Pacemakers
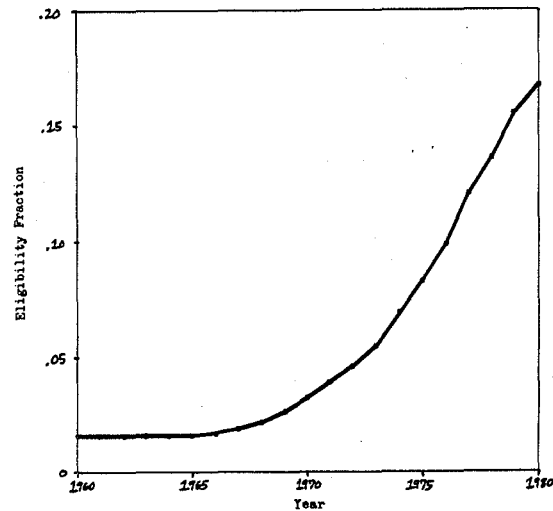(U.S., 1960-1980) Source: Homer 1983 ([8])

Figure 2. Fraction of Arrhythmic Patients Deemed
Eligible for Pacing
Source: Homer 1983 ([8])

9

Journal articles are a source of information which can affect the
extent to which a new medical technology is used.[10][11][12][13] The
articles which have the greatest impact are those which appear in influ-
ential journals, describe clinical results in detail, and report on a
large sample of cases. Several physicians, when interviewed, pointed to
four journals which they consider to be most informative on the subject
of pacing.[14] These journals were searched thoroughly for articles which
appeared during the period 1960-1980 and which presented clinical follow-up
data on the observed results of pacing for specified medical conditions.
Several different kinds of information were gleaned from the fifty-six arti-
cles which satisfied these criteria, including the number of recipients
reported per article. This information was used to construct Figure 3,
which shows the total cases reported in the four journals on an annual ba-
sis (the "reporting rate"). Since larger sample sizes have greater impact,
this time series may be considered to be an indication of how the influence
of pacing articles changed over time.

The observed pattern of reporting is clearly oscillatory and appears
to have a period of about five years, with an amplitude that is small at
first but then becomes larger. This time series is striking, particularly
because it is the only one of several pacing time series that were compiled
that displays oscillations. Such oscillations are nowhere discussed in the
literature. Indeed, very little has been written on the factors that under-
lie evaluation and publication decisions in general.[15] The component of
the model in which the reporting rate is generated therefore seems to be a
prime candidate for partial-model testing: Can this component (whose struc-
ture is largely speculative) reproduce the observed oscillations, given
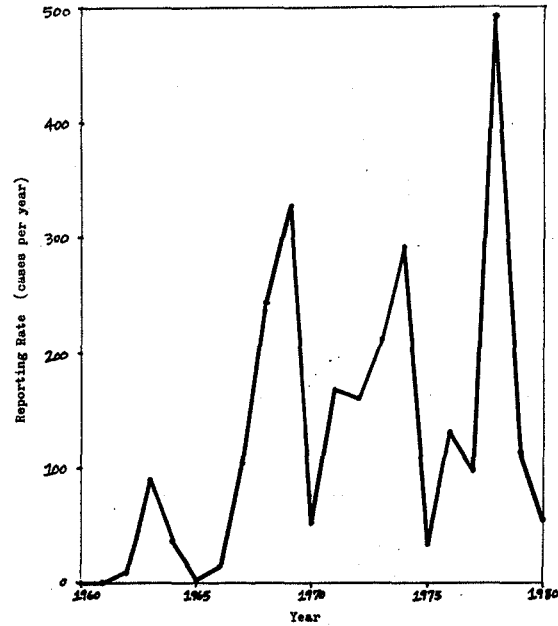inputs which are not themselves oscillatory?

Figure 3. Paced Patients Reported in Influential Journals
Source: Homer 1983 ([8])

3.2 Evaluation and Reporting Subsystem

Since little has been written about the processes of evaluation
and reporting, the structure to be presented here is based primarily on
inferences drawn from the case studies (particularly the pacemaker case
study).[16] The goal was to make this structure as simple and transparent
as possible by making only a small number of common-sense assumptions which
seem generalizable beyond the specific context of the case studies. The
resulting structure is shown in Figure 4. (A DYNAMO listing of the evalu-
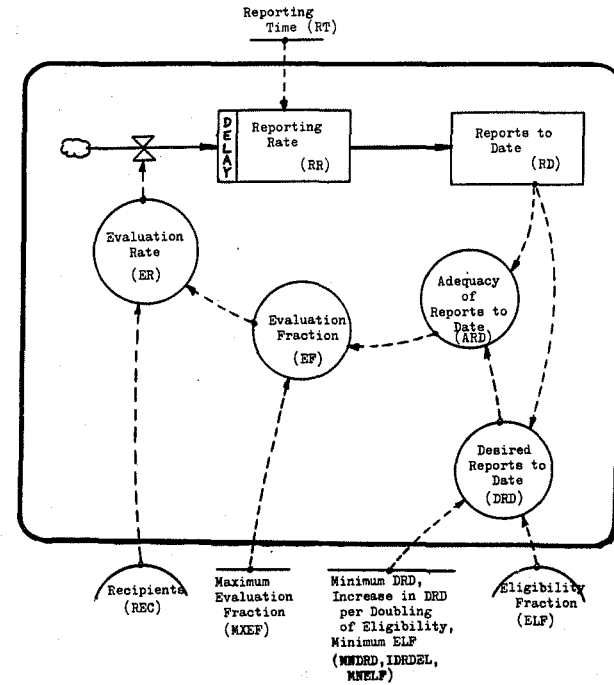ation and reporting subsystem can be found in this paper's Appendix.)

Figure 4. Evaluation and Reporting Subsystem

The reporting rate (RR), as discussed previously, represents the
publication in influential journals of clinical case information and is
measured in cases published per year. Thus, reports to date (RD) repre-
sents the cumulative number of recipients whose procedures appear in the
broadly recognized clinical literature on the technology. The reporting
rate is simply a delayed version of the evaluation rate (ER), which corre-
sponds to the annual number of cases selected for analysis which do eventu-
ally make it into the journals.[17] The delay separating evaluation and

reporting represents the total time required to write, submit, and publish an article on clinical outcomes. A comparison of submission and publication dates for pacemaker reports suggested that the reporting time (RT) generally exceeds one-half year and may even exceed two years, although one year is more typical. (This parameter can be estimated using partial-model tests, as discussed below.)

Physicians select patients for evaluation from the total pool of recipients, in response to a perceived need for more evaluative data on the technology. Both evaluators and journal editors will tend to turn their attentions to other topics as they become more certain of the technology's capabilities and limits. In other words, the fraction of recipients selected for evaluation (the evaluation fraction, EF) will be responsive to the adequacy of existing reports (adequacy of reports to date, ARD); the fraction will be greatest (equal to the maximum evaluation fraction, MXEF) when no data are available and smallest when reports are seen as fully adequate. This relationship is represented analytically in the model as $EF=(MXEF)(1-ARD)^2$, a nonlinear function of ARD, which says that evaluators will be more sensitive to changes in the need for data when adequacy is low than when it is high. One can think of this curve as reflecting the relative number of physicians who choose to investigate the technology in response to a given level of need: When the adequacy of reports is high, only a small number of researchers with special commitment will continue to evaluate the technology, while the bulk of physicians will only enter the fray if uncertainty is high and the journals are especially anxious to print new data on the technology. In the pacemaker case, a national survey showed that fully a third of implanting physicians have published articles on pacing,[18] although prob-

ably most of them no more than once or twice. On the basis of this information, the maximum evaluation fraction (MXEF) was set at 0.3 (30% per year), which says that virtually all of these physicians would respond in the event that no data on pacing existed.

The adequacy of reports (ARD) is simply the ratio of existing reports (RD) to the desired level of reports to date (DRD). This desired level will be no smaller than the current stock of information, but may be expected to increase as the eligibility fraction (ELF) increases. One may look at this from a statistical perspective: As eligibility expands, the population being assessed grows proportionally, so that a larger sample is needed to achieve a given level of confidence in the results. More concretely, eligibility criteria widen through the inclusion of new subsets of patients, and since outcomes may be highly dependent on the particular subset being considered,[12][19][20][21] such an expansion in scope generally implies a need for new data. The pacemaker case provides a good example of this phenomenon: In the early 1960s, articles focused on results for patients who had severe symptoms and were at high risk of sudden death; in the late 1960s, the focus switched to a newly-included subset of patients with less risk of sudden death; in the early 1970s, articles concentrated on patients with "sick sinus syndrome", whose risk of sudden death is negligible; and in the late 1970s, the focus of articles shifted to the benefits of the "prophylactic" pacing of patients who have few or no symptoms associated with their arrhythmias but are at some risk of sudden death (such as certain heart attack victims).[22] Indeed, one can largely associate each fluctuation in Figure 3 with a period of renewed clinical research activity brought on by the application of pacing to a new subset (or subsets) of arrhythmic patients.

14

The following nonlinear relationship was selected to describe the connection between the eligibility fraction and the desired reports to date;[23]

$$DRD = Max\ (RD,\ MNDRD\ +\ IDRDEL*log_2(ELF/MNELF))$$

where   MNDRD is the minimum desired reports to date
       IDRDEL is the increase in desired reports to date
        per doubling of eligibility
       MNELF is the minimum eligibility fraction

This formulation was suggested both by theoretical considerations (sampling theory suggests the required sample size should rise as the square root of the population under consideration) and by a comparison of the historical values of reports to date (RD) and the eligibility fraction (ELF) in the pacemaker case. This comparison is shown in Figure 5, graphed in semi-log form. Each point on the graph corresponds to one year between 1960 and 1980.[24] If one assumes that the adequacy of reports was for the most part high during this period--that is, that the response to needed data has always been relatively rapid and unhindered (this description does seem to fit the case of pacing well)--then reports to date can be used as a rough proxy for _desired_ reports to date, and the observed relationship can be used to estimate the two parameters MNDRD and IDRDEL. The straight line superimposed on the empirical data in Figure 5 shows that a good fit is obtained when MNDRD=100 cases and IDRDEL=700 cases. (Note that both estimates pass the most basic common-sense criterion, namely, that they both be positive.)

Since aggregate time-series data were used to derive these two estimates, what has just been described is actually an example of partial-
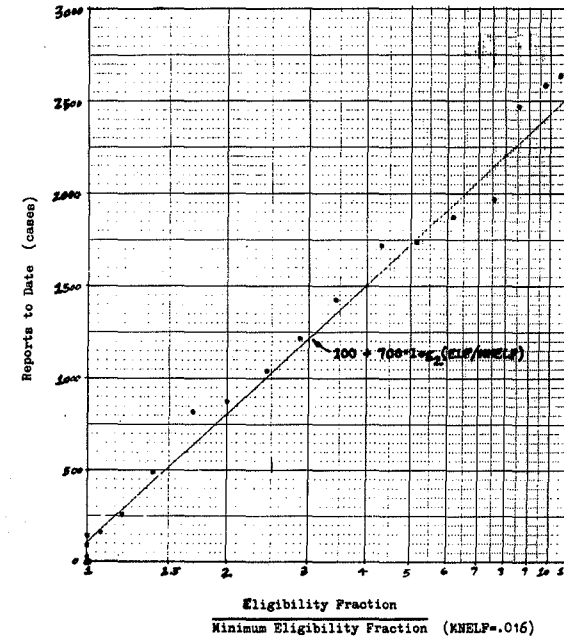
15

Figure 5.   Eligibility Fraction versus Reports to Date

model parameter estimation (even if the piece of structure is only a single equation). Thus, some partial-model testing of a small piece of the evaluation and reporting subsystem was found to be necessary before the whole subsystem could be tested. In general, the strategy of testing components which are as small as the data will permit can lead to such a _sequence_ of partial-model tests, in which tests of one piece of the model are best performed after another uncertain piece has been tested and adjusted. In the present instance, the evaluation and reporting loop can now be tested without further adjustment of the parameters MNDRD and IDRDEL.

16

3.3 Simulation Testing

In this section, the use of partial-model testing to estimate the order and length of the delay between evaluation and reporting will be demonstrated.[25] The time series on recipients and eligibility fraction, shown in Figures 1 and 2, respectively, were used as inputs in this testing, and the simulated reporting rate compared with the historical data shown in Figure 3. In the full medical technology model, both recipients and the eligibility fraction are generated endogenously, and the reporting rate can feed back, through various channels, to affect these variables. Because of this feedback, a successful partial-model test does not guarantee equal success in the context of the full model. But as discussed previously, this is precisely the sacrifice made in exchange for the greater structural validity gained when structures are estimated using the smallest possible number of assumptions and aggregate time series in any given test.

In the interest of simplicity of structure, a first-order delay (DELAY1) was initially used for testing purposes. The reporting time was adjusted between one-half and two years (the range suggested by prior information) until a simulated reporting rate with period and amplitude characteristics most like those of the historical time series was found. Figure 5 shows the best result, obtained when the reporting time was set to 1.25 years. The simulated fluctuations are neither as strong nor as regular as those seen in the empirical data. An astute system dynamicist might guess that the first-order delay does not introduce enough of a phase lag into the subsystem's negative loop to bring out this loop's oscillatory potential to a sufficient degree. Increasing the order of the delay therefore makes sense as the next step in partial-model testing. But in the present instance,
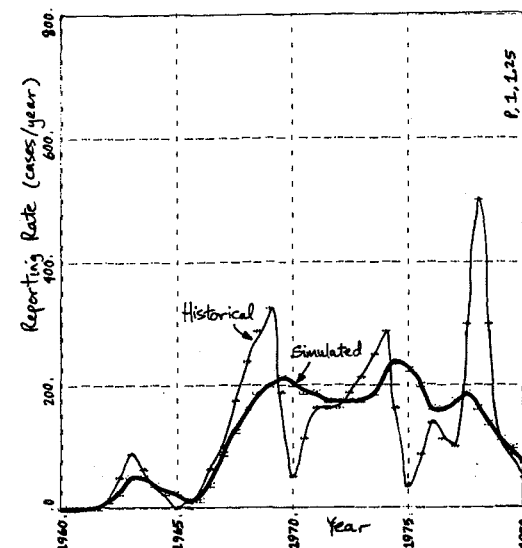
17

Figure 6. Partial-Model Test, First-Order Delay

one need not argue for a higher-order delay (DELAY3 instead of DELAY1) on these behavioral grounds alone; the fact that the "reporting delay" actually represents the multi-stage process of writing, submitting, and publishing an article means that a structural defense for the implied increase in model complexity exists as well. (Indeed, rarely will a physician be in the position of evaluating patients one day and seeing his results published the next, as would happen frequently if the process were really a first order one.)

The move to a third-order delay does, in fact, improve the subsystem's response considerably. Figure 7 snows the best result, again obtained when the reporting time was set to 1.25 years. Not only are the frequency and
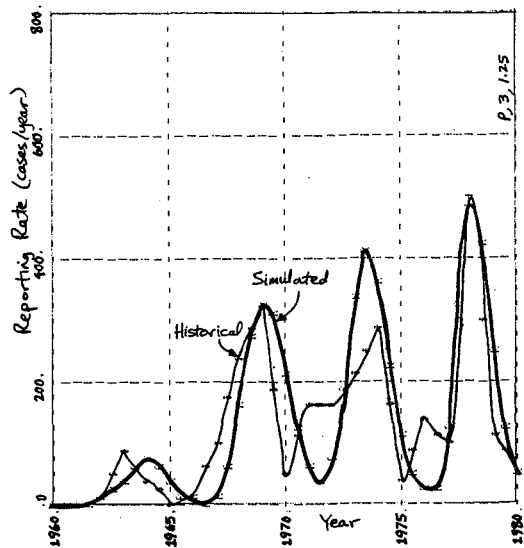
18

Figure 7. Partial-Model Test, Third-Order Delay

(growing) amplitude of the simulated oscillation true to the original, but
the timing matches as well. Surely this test serves to strengthen one's
confidence in the evaluation and reporting subsystem!

At this juncture, it is appropriate to explain the observed oscil-
lations from the vantage-point of the model. The bottom line here is that
the reporting rate can grow to exceed that rate required to just satisfy
the need created by expanding eligibility. The source of this information
glut is the delay between evaluation and reporting, which enables a backlog
of not-yet-published evaluations to accumulate. If the publication process
were fully centralized, this would not cause a problem, because the backlog
would be taken into account. But since there are many evaluators and gener-

19

ally several journals to which they can send evaluations, several reports
on the same subject may appear simultaneously even if only one of them is
necessary to satisfy the current need for information.

Given this background on how over-reporting can occur, the follow-
ing story can be told for the pacemaker case: As the eligibility fraction
increased exponentially over time, the demand for reports also increased
(roughly linearly). In response to the increasing demand for evaluative
data, reports were submitted and published at a rate that turned out to
somewhat exceed current needs and which led to periods during which the
cumulative stock of reports appeared adequate. But the eligibility frac-
tion continued to climb, so the complacency would eventually wear off and
a whole new round of evaluation and reporting would begin. As the number
of recipients who could be evaluated grew, the response to a given need
for data increased, which explains the growing amplitude of oscillation.

It is interesting to examine how sensitive the simulated oscilla-
tion is to the delay time. Figure 8 shows the result of reducing the delay
time to one-half year, while Figure 9 shows the result of increasing the
delay time to two years. Clearly, as the delay time increases, both the
period and amplitude of oscillation also increase. The longer the report-
ing delay is, the greater the potential for excess evaluation and reporting
activity, so the greater the amount of overshoot and the longer it takes for
the resulting complacency to be worn away by expanding eligibility. From
these and other partial-model tests, it appears that the period of oscil-
lation is roughly proportional to (about four times) the reporting delay
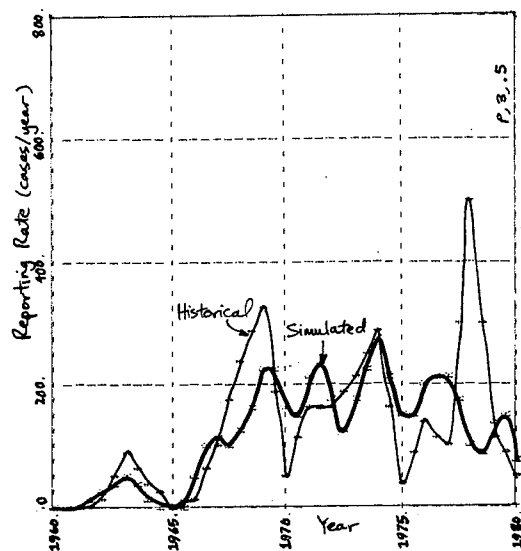time, a high degree of sensitivity by any measure.

20

Figure 8. Partial-Model Test, Third-Order Delay,
Shorter Delay Time

21

Figure 9. Partial-Model Test, Third-Order Delay,
Longer Delay Time

Having settled upon formulations and parameter values which perform well in the partial-model context, it is desirable to examine the simulated reporting rate in the context of the full model. Figure 10 presents a comparison of whole-model simulation results with the empirical data. Even with all of the model's feedback loops intact—specifically, with recipients and eligibility fraction being generated endogenously—the amplitude and period of oscillation are still reproduced faithfully. However, this simulation does show that the timing of oscillation is affected; the simulated reporting rate lags the historical rate by about a year throughout the twenty-year period. Apparently, the slight differences between the
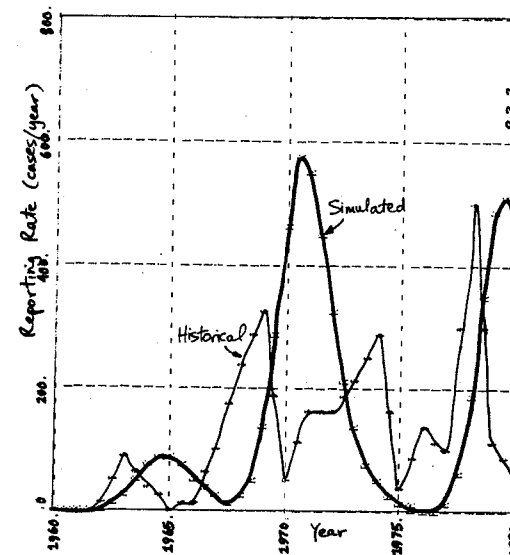
simulated and historical values of recipients and eligibility (and these differences are quite small) are enough so that the oscillations are triggered at slightly different times in the whole-model test than in the partial-model test.

This raises an interesting point. Since it has been shown that the evaluation and reporting subsystem can produce the correct timing of oscillations under the controlled conditions of a partial-model test, one is led to the conclusion not that something is wrong with the model, but rather that, in real life, the precise timing of the oscillations is sensitive to random disturbances that can affect the evaluation and reporting process.
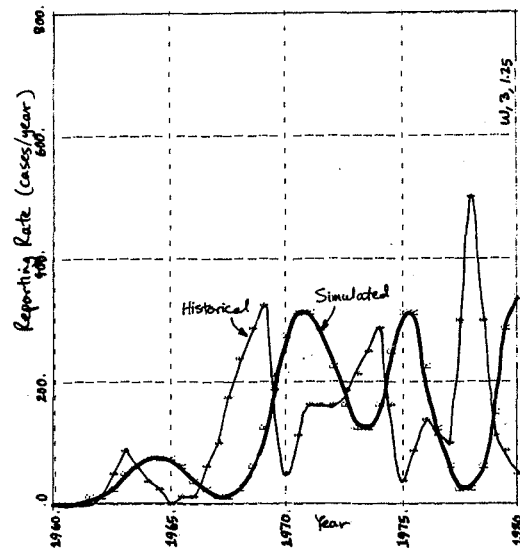
22

Figure 10. Whole-Model Simulation of the
Reporting Rate

Thus, although the period and amplitude of oscillation appear to be
robust features of the system's behavior, timing seems to be a more unpre-
dictable matter. It would be difficult to make this point convincingly
without performing a partial-model test that reproduces the historical
timing of oscillation.

4. Conclusion

This paper has outlined and demonstrated an approach to pinning
down uncertain formulations and parameter values. In actual modeling
practice, one frequently does not have access to as much disaggregate

---

23

case-specific or logical information as would be necessary to specify a
complex model in its entirety. In such circumstances, some comprimising
of the ideal a priori approach to model construction is necessary, leading
to a certain amount of curve-fitting. This paper has argued that there
are better and worse ways of going about this, given the desire to avoid
"fixing" the final result. The key is doing partial-model testing instead
of whole-model testing whenever possible. More generally, the suggested
strategy is to set formulations and parameter values on the basis of tests
of model components which are as small as the available time series data
permit. This approach not only minimizes the likelihood of incorrectly
attributing behavior to "innocent" (or only marginally involved) relation-
ships or parameters, but can also lend greater efficiency to the process
by which a model is improved.

NOTES AND REFERENCES

[1]   Forrester JW. Industrial Dynamics. MIT Press, Cambridge: 1961;
       pp. 115-129.

[2]   Forrester JW. Confidence in Models of Social Behavior—with Empha-
       sis on System Dynamics Models. MIT System Dynamics Memo D-1967,
       1973.

[3]   Mass NJ. "Diagnosing Surprise Model Behavior: A Tool for Evolving
       Behavioral and Policy Insights," presented at the 1981 System
       Dynamics Research Conference, Rensselaerville, New York: 1981.

[4]   Graham AL. "Parameter Formulation and Estimation in System Dynamics
       Models," in The System Dynamics Method: The Proceedings of the
       1976 International Conference on System Dynamics, ed. J Randers,
       LK Ervik. Resource Policy Group, Oslo, Norway: 1977.

[5]   Hamilton MS. "Estimating Lengths and Orders of Delays in System
       Dynamics Models," in The System Dynamics Method, Oslo: 1977.
       This paper explores a variety of statistical estimation methods
       using input and output data.

24

[5] Naill RF. "The Discovery Life Cycle of a Finite Resource: A Case Study of U.S. Natural Gas," in Toward Global Equilibrium, ed. DL Meadows, DH Meadows. Wright-Allen Press, Inc., Cambridge: 1973. This paper uses a statistical method to estimate the delay time separating natural gas explorations and discoveries.

[7] Peterson DW. "Statistical Tools for System Dynamics," in The System Dynamics Method, Oslo: 1977.

[8] Homer JB. A Dynamic Model for Analyzing the Emergence of New Medical Technologies. PhD Dissertation, MIT: 1983.

[9] See Chapter 3 of [8] on pacing.

[10] Stross JK, Harlan WR. The dissemination of new medical information. Journal of the American Medical Association, 241: 2622-2624, 1979.

[11] Manning PR, Denson TA. How internists learned about cimetidine. Annals of Internal Medicine, 92: 690-692, 1980.

[12] Banta HD, Behney CJ, Willems JS. Toward Rational Technology in Medicine. Springer Publishing Co., New York: 1981.

[13] Young DA. "Communications Linking Clinical Research and Clinical Practice," in Biomedical Innovation, ed. EB Roberts and others. MIT Press, Cambridge: 1981.

[14] The four journals are: 1) PACE (Pacing and Clinical Electrophysiology), 2) Circulation, 3) The American Journal of Cardiology, 4) The American Heart Journal. See Appendix P3 of [8].

[15] On a related topic: Levy RI, Sondik EJ. Decision-making in planning large comparative studies. Annals of the NY Academy of Sciences, 304: 441-457, 1978.

[16] The structure presented here is a bit simpler than that used in [8], which includes a "controversy" factor which was not of importance in the pacemaker case.

[17] Note the assumption that all evaluations are eventually published, which seems to ignore the process by which papers are screened for publication and many rejected. In fact, the screening process is represented in the model, but for the sake of simplicity, has been placed at the stage of evaluation rather than reporting.

[18] Parsonnet V, Manhardt M. Permanent pacing of the heart: 1952 to 1976. American Journal of Cardiology, 39: 250-256, 1977.

[19] Weinstein MC, Stason WB. Foundations of cost-effectiveness analysis for health and medical practices. New England Journal of Medicine, 296: 716-721, 1977.

25

[20] Fineberg HV, Hiatt HH. Evaluation of medical practices: The case for technology assessment. New England Journal of Medicine, 301: 1086-1091, 1979.

[21] National Academy of Sciences. "The Evaluation of Equipment-Embodied Technology," in A Study of the Diffusion of Equipment-Embodied Technology. Government Printing Office, Washington DC: 1978.

[22] See Appendix P3 of [8] on the pacemaker literature.

[23] DRD is never less than RD and has an absolute minimum value of MNDRD, which corresponds to a situation in which ELF=MNELF. In the pacemaker case, the eligibility fraction (ELF) was equal to its minimum value (MNELF) of .016 during the period 1960-1965; physicians were extremely cautious in their selection of patients to receive the device during these early years.

[24] The eligibility fraction series was plotted in Figure 2. The reports to date series was calculated by accumulating the reporting rate series, which was plotted in Figure 3.

[25] Hamilton ([5]) discusses various single-equation estimation techniques for this problem when data on both the direct input to the delay (the evaluation rate) and the output to the delay (the reporting rate) are available. But since data on the evaluation rate were not available, a larger piece of structure—in fact the whole evaluation and reporting subsystem—was required for estimation purposes.

APPENDIX: Dynamo Program Used for Partial-Model Testing

```
*     EVALUATION AND REPORTING SUBSYSTEM, PACEMAKER CASE
A     RR.K=DELAY3(ER.JK,RT)                                        1
C     RT=1.25 years
R     ER.KL=(REC.K)(EF.K)                                          2
A     EF.K=(MXEF)(1-ARD.K)(1-ARD.K)                                3
C     MXEF=.3  per year
A     ARD.K=RD.K/DRD.K                                             4
L     RD.K=RD.J+(DT)(RR.J)                                         5
N     RD=0
A     DRD.K=MAX(RD.K,MNDRD+(IDRDEL*1.443*LOGN(ELF.K/MNELF)))       6
C     MNDRD=100 cases
C     IDRDEL=700 cases
C     MNELF=.016
NOTE   TIME SERIES FOR PARTIAL-MODEL TESTS
A     REC.K=1E3*TABLE(TREC,TIME.K,1960,1980,1)                     7
T     TREC=0/0/0.2/0.7/1.3/2.1/3.0/4.3/7.2/13.5/23.6/36.8/
X       52.7/71.2/93.8/121.6/155.2/195.8/243.2/296.3/354.6
A     ELF.K=TABLE(TELF,TIME.K,1960,1980,1)                         8
T     TELF=.016/.016/.016/.016/.016/.016/.017/.019/.022/
X       .027/.032/.039/.046/.055/.069/.083/.099/.121/.137/
X       .156/.168
A     HRR.K=TABLE(THRR,TIME.K,1960,1980,1)                         9
T     THRR=0/0/7/89/38/2/15/103/242/324/54/168/160/212/
X       291/35/132/98/495/116/55
NOTE   RUN STATEMENTS
SPEC   DT=.125/LENGTH=1980/PLTPER=.5
N      TIME=1960
PLOT   RR=R,HRR=$(0,800)
RUN    FIGURE 7
```