Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

CREDIBILITY ASSESSMENT OF SIMULATION RESULTS†

Osman Balci

Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

ABSTRACT

The purpose of this paper is to provide some guidelines for assessing the credibility of simulation results. The life cycle of a simulation study is characterized in terms of 10 phases, 10 processes, and 13 credibility assessment stages (CASs). The credibility of simulation results is assessed by integrating ten CASs: formulated problem verification, feasibility assessment, system and objectives definition verification, model qualification, communicative model verification, programmed model verification, experiment design verification, data validation, model validation, and quality assurance of experimental model. Indicators are identified for evaluating credibility in most of the CASs. The guidelines provided herein are essential for the success of a simulation study.

1. INTRODUCTION

Since the advent of computers, simulation has emerged as one of the most powerful techniques for problem solving in many disciplines. As the problems grow larger, become more complex, and require more precise solutions than ever before, the use of simulation becomes more frequent. As reported by Roth et al. (1978), the U.S. Government is the largest sponsor and consumer of models in the world. Estimates have indicated that over one-half billion dollars are being spent annually on developing, using, and maintaining mathematical, simulation, and econometric models in the decision-and-policy-making functions of the Federal Government.

In a simulation study, we work with a model of the problem rather than directly working with the problem itself. If the model does not possess a sufficiently accurate representation, we can easily have "junk input" and "junk output." Simulation model development should not be viewed as strictly software development and the art of modeling should be mastered. It is no challenge to write a computer program which accepts a set of inputs and produces a set of outputs to do simulation. The challenge is to do it right. Multifaceted and multidisciplinary knowledge and experience are required for a successful simulation study.

The purpose of this paper is to present some guidelines for assessing the credibility of simulation results. The paper begins by introducing the life cycle of a simulation study. This is followed by describing the ten CASs and finally some concluding remarks are given.

2. THE LIFE CYCLE OF A SIMULATION STUDY

The life cycle is presented in Figure 1 [reprinted from (Balci 1986b)]. Oval symbols represent the phases. The dashed arrows describe the processes which relate the phases to each other. The solid arrows refer to the credibility assessment stages (CASs).

The life cycle should not be interpreted as strictly sequential. The sequential representation of the dashed arrows is intended to show the direction of development throughout the life cycle. The life cycle is iterative in nature and reverse transitions are expected.

The phases starting with System and Objectives Definition and culminating with Model Results correspond to model development phases. A Model Management System (Nance and Balci 1986; Nance et al. 1981) covers the entire life cycle and Model Development Environments (Balci 1986a; Nance 1983) cover the model development phases.

In the life cycle in Figure 1: Input Data Modeling is a subprocess of Model Formulation; Simulation Programming Languages, Random Number Generation, Random Variate Generation, and Time Flow Mechanisms are subprocesses of programming; and Statistical Analysis of Simulation Output Data/Design of Simulation Experiments are subprocesses of Design of Experiments and Experimentation.

3. CREDIBILITY ASSESSMENT

In Elmaghraby's words, "It is well to remember the dictum that nobody solves *the* problem. Rather, everybody solves the model that he [or she] has constructed of the problem" (Elmaghraby 1968, p. 305). Thus it is crucial that we assess the credibility of each process as we progress in the life cycle.

Since a model is an abstraction of the reality, we cannot talk about its *absolute* accuracy. Credibility, quality,

† This paper is extracted from O. Balci "Guidelines for Successful Simulation Studies," Technical Report TR-85-2, Department of Computer Science, Virginia Tech, Blacksburg, Va., May 1986, 55 pp.

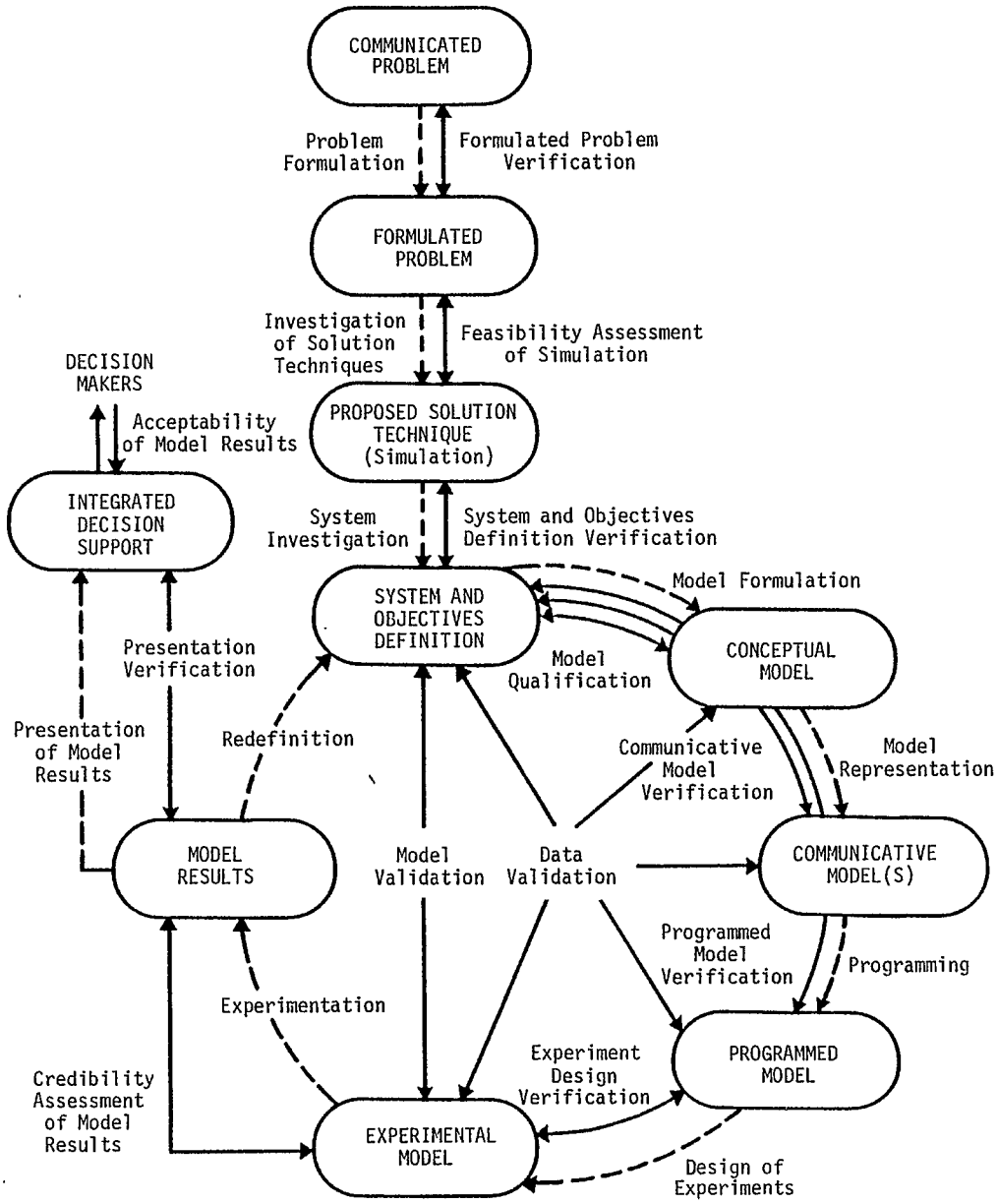

Figure 1: The Life Cycle of a Simulation Study.

validity, and verity are measures that are assessed with respect to the study objectives for which the model is intended. In some cases, a 60% level of confidence in the credibility of model results may very well serve for the purpose; in another, 90% may be required.

Three types of errors may be committed in conducting a simulation study. *Type I Error* is committed when the study results are rejected when in fact they are sufficiently credible. *Type II Error* is committed when the study results are accepted when in fact they are *not* sufficiently credible. The definitions of type I and type II errors can be extended to apply for every CAS. In the case of model validation, we called the probability of committing type I error as *model builder's risk* and the probability of committing type II error as *model user's risk* (Balci and Sargent 1981). *Type III Error* is committed when the formulated problem does not completely contain the *actual* problem (Balci and Nance 1985). Committing type III error corresponds to solving the wrong problem.

A simulation project team should possess at least four areas of knowledge and experience to be successful: (1) project leadership, (2) modeling, (3) programming, and (4)

knowledge of the system under study (Annino and Russell 1979). Lacking the necessary level of expertise in any area may result in a failure of the project or an error of type I, II or III.

A subjective, yet quite effective method for evaluating the acceptability of model results is *peer assessment*, the assessment of the acceptability by a panel of expert peers. This panel should be composed of (1) people who have expert knowledge of the system under study, (2) expert modelers, (3) expert simulation analysts, and (4) people with extensive experience with simulation projects.

The panel examines the overall study based upon the project team's presentation and detailed study of documentation. Working together and sharing their knowledge among each other, panel members measure the indicators identified for the CASs of the life cycle. An indicator is an indirect measure of a concept and it can be measured directly (Balci 1986b).

An indicator is weighted and measured with a score [See (Balci 1986b) for details]. The higher the overall score the more confidence we gain for the acceptability of model results. However, even a perfect score would not guarantee that the results will be accepted and used by the decision makers; because, acceptability is an attribute of the decision maker not an attribute of the simulation study. Perfect results may be rejected due to the lack of credibility of the institution performing the study or due to a political reason. Nevertheless, the objective of the simulation project management should be to increase the confidence as much as possible. A higher overall score may not guarantee the acceptance of results but a lower overall score can easily result in their rejection or an error of type II.

3.1 Formulated Problem Verification

"Substantiation that the formulated problem contains the *actual* problem in its entirety and is sufficiently well structured to permit the derivation of a sufficiently credible solution" is called *formulated problem verification* (Balci and Nance 1985). For this substantiation, a Questionnaire is developed in (Balci and Nance 1985) with 38 indicators. People who are intimately knowledgeable of the problem(s) based on experience and training verify the formulated problem by measuring the indicators. The reader is referred to (Balci and Nance 1985) for the details of the verification.

3.2 Feasibility Assessment of Simulation

Are the benefits and cost of simulation solution estimated correctly? Do the potential benefits of simulation solution justify the estimated cost of obtaining it? Is it possible to solve the problem using simulation within the time limit specified? Can all of the resources required by the simulation project be secured? Can all of the specific requirements (e.g., access to pertinent classified information) of the simulation project be satisfied? These questions are the indicators of the feasibility of simulation.

3.3 System and Objectives Definition Verification

We should justify that the system characteristics are identified and the study objectives are explicitly defined with sufficient accuracy. An error made here may not be caught until very late in the life cycle resulting in a high cost of correction or an error of type II or III.

Since systems and objectives may change over a period of time, will we have the same system and objectives definition at the conclusion of the simulation study (which may last from 6 months to several years)? Is the system's environment (boundary) identified correctly? What counterintuitive behavior may be caused within the system and its environment? Will the system significantly drift to low performance requiring a periodic update of the system definition? Are the interdependency and organization of the system characterized accurately?

3.4 Model Qualification

A model should be conceptualized under the guidance of a structured approach such as the Conical Methodology (Nance 1981) which is a top-down definition and a bottom-up specification approach for building models. One key idea behind the use of a structured approach is to control the model complexity so that we can successfully verify and validate the model. The use of a structured approach is an important factor determining the success of a simulation project, especially for large-scale and complex models.

During the conceptualization of the model, one makes many assumptions in abstracting the reality. Each assumption should be explicitly specified. Model Qualification deals with the justification that all assumptions made are appropriate and the conceptual model provides an adequate representation of the system with respect to the study objectives.

3.5 Communicative Model Verification

In this stage, we confirm the adequacy of the communicative model to provide an acceptable level of agreement for the domain of intended application. *Domain of Intended Application* (Schlesinger et al. 1979) is the prescribed conditions for which the model is intended to match the system under study. *Level of Agreement* (Schlesinger et al. 1979) is the required correspondence between the model and the system, consistent with the domain of intended application and the study objectives.

Communicative Model Verification (CMV) and Programmed Model Verification (PMV) can be achieved by using one or more testing techniques that are summarized in Table 1.

Table 1: A summary of testing techniques for the verification of communicative and programmed models.

Technique	Manual/ Automated	Static/ Dynamic	Structural/ Functional	Usable for CMV/PMV
Desk Checking	manual	both	structural	both
Model Review	manual	both	structural	both
Graph-Based Analysis	automated	static	structural	both
Instrumentation-Based Testing	automated	dynamic	structural	PMV
Functional Testing	both	dynamic	functional	PMV

A dynamic analysis technique requires that the model be executed, whereas static analysis does not. If testing can be applied with mental execution, it is considered dynamic. Structural testing deals with the analysis of model's internal structure, while functional testing is concerned with model's input-output transformation. See (Balci 1986b) for descriptions of these techniques.

3.6 Programmed Model Verification

Substantiation that the programmed model has sufficient amount of accuracy in representing the communicative model is called PMV. All testing techniques summarized in Table 1 can be used for PMV. See (Balci 1986b) for descriptions of these techniques.

3.7 Experiment Design Verification

The design of experiments can be verified by measuring the following indicators: (1) Are the algorithms used for random variate generation theoretically accurate?; (2) Are the random variate generation algorithms translated into executable code accurately? (Error may be induced by computer arithmetic (Monahan 1977) or by truncation due to machine accuracy, especially with order statistics (e.g., $X = -\log_e(1-U)$) (Schmeiser 1981).); (3) How well is the random number generator tested? (Using a generator which is not rigorously shown to produce uniformly distributed independent numbers with sufficiently large period may invalidate the whole experiment design.); (4) Are *appropriate* statistical techniques implemented to design and analyze the simulation experiments? How well are the underlying assumptions satisfied? (See (Law 1983) for several reasons why output data analyses have not been conducted in an appropriate manner.); (5) Is the problem of the initial transient (or the start-up problem) (Wilson and Pritsker 1978) appropriately addressed?; and (6) For comparison studies, are identical experimental conditions replicated correctly for each of the alternative operating policies compared?

3.8 Data Validation

In this stage, we confirm that the data used throughout the model development phases are accurate, complete, unbiased, and appropriate in their original and transformed forms. The data used can be classified as (1) model input data and (2) model parameters data.

Data validation deals with the substantiation that each input data model used possesses satisfactory accuracy consistent with the study objectives, and that the parameter values are accurately identified and used.

Here are some indicators to measure data validity: (1) Does each input data model possess a sufficiently accurate representation?; (2) Are the parameter values identified, measured, or estimated with sufficient accuracy?; (3) How reliable are the instruments used for data collection and measurement?; (4) Are all data transformations done accurately? (e.g., are all data transformed correctly into the same time unit of the model?); (5) Is the dependence between the input variables, if any, represented by the input data model(s) with sufficient accuracy? (Blindly modeling bivariate relationships using only correlation to measure dependence is cited as a common error by Schmeiser (1981).); and (6) Are all data up-to-date?

3.9 Model Validation

Substantiating that the experimental model, within its domain of applicability, behaves with satisfactory accuracy consistent with the study objectives is called *Model Validation.* The *Domain of Applicability* is the set of prescribed conditions for which the experimental model has been tested, compared against the system to the extent possible, and judged suitable for use (Schlesinger et al. 1979).

Model validation is performed by comparing model behavior with system behavior when both model and system are driven under identical input conditions. Only under those input conditions we can claim model validity, because a model which is sufficiently valid under one set of input conditions can be completely absurd under another (Zeigler

1976). If a model is used in a "what if" environment or if it is a forecasting model, the possible input conditions may form a very large domain over which model validation may become infeasible.

The existing literature on simulation model validation (Balci and Sargent 1984) generally falls into two broad areas: subjective validation techniques and statistical techniques proposed for validation. See (Balci 1986b) for descriptions of these techniques.

3.10 Quality Assurance of Experimental Model

There are other indicators, described below, in addition to the ones presented in Sections 3.4—3.9 for assuring the quality of experimental model. These indicators are derived from software quality characteristics (Boehm et al. 1976).

(1) *Accessibility:* Does the model facilitate selective use of its parts for other purposes (e.g., for the construction of another model)?

(2) *Accountability:* Does the model lend itself to measurement of its usage? Can probes be inserted to measure timing, whether specified branches are exercised, etc.?

(3) *Accuracy:* Are the model's calculations and outputs sufficiently precise to satisfy their intended use?

(4) *Augmentability:* Can the model accommodate expansion in component computational functions or data storage requirements?

(5) *Communicativeness:* Does the model facilitate the specification of inputs? Does it provide outputs whose form and content are easy to assimilate and useful?

(6) *Completeness:* Are all model inputs used within the model? Are there no "dummy" submodels referenced?

(7) *Conciseness:* Is the model implemented with a minimum amount of code? Is it excessively fragmented into submodels so that the same sequence of code is not repeated in numerous places?

(8) *Consistency:* Does the model contain uniform notation, terminology, and symbology within itself? Are all model attributes and variables typed and specified consistently for all uses? Are coding standards homogeneously adhered to?

(9) *Device-independence:* Can the model be executed on other computer hardware configurations? Have machine-dependent statements been flagged and documented?

(10) *Efficiency:* Does the model fulfill its objective without waste of resources?

(11) *Legibility:* Does the model possess the characteristic that its function is easily discerned by reading the code?

(12) *Self-containedness:* Does the model perform everything needed for its execution within itself? Does it not require to use a database, a library of routines or an application program?

(13) *Self-descriptiveness:* Does the model contain enough information for a reader to determine or verify its objectives, assumptions, constraints, inputs, outputs, components, and revision status?

(14) *Structuredness:* Does the model possess a definite pattern of organization of its interdependent parts?

(15) *Robustness:* Does the model continue to execute reasonably when it is run with invalid inputs? Can the model assign default values to non-specified input variables and parameters? Does the model have the capability to check input data for domain errors?

3.11 Credibility Assessment of Simulation Results

Concluding upon sufficient quality of the experimental model is a necessary but not a sufficient requirement for the credibility of model results. (The term "model results" is preferred over "simulation results" so as to emphasize that the results are model-based.) The experimental model quality is assured with respect to the definition of system and study objectives. An error made in defining the system or a study objective or failing to identify the *real* problem may cause unacceptable model results or an error of type II. The credibility of model results is assessed by way of performing the ten CASs discussed in Sections 3.1 to 3.10.

4. CONCLUDING REMARKS

Although the life cycle of a simulation study is characterized with 13 CASs, a literature review (Balci and Sargent 1984) reveals that most work has concentrated on model validation and very little has been published on the other CASs. Model validity is a necessary but not a sufficient requirement for the credibility of model results. Sufficient attention must be devoted to every CAS in order for a simulation study to be successful.

A simulation study is multifaceted and multidisciplinary as illustrated by the life cycle presented herein. Ören in his forward to (Zeigler 1984) indicates that

> "Some of the specialists too close to one of the facets, perceive only that single facet and the reflection of the success of their careers through it. The more they see the latter, the more, it seems, they are enamoured with that aspect of modeling and simulation instead of exploring new horizons. If it was to this attitude, nobody would have discovered the New World."

Sufficient effort must be devoted to every process of the life cycle. Inadequate coverage of a process, due to insufficient knowledge or time, may result in unacceptable results or an error of type II.

The list of indicators for the CASs is not intended to be exhaustive. Additional indicators that are specific to the area of application should be employed whenever possible. There is also the issue of assessing the validity and reliability of these indicators which is extremely difficult if not impossible for the broad scope adopted herein; however, for a specific area of application this should be achievable.

ACKNOWLEDGMENTS

This research was sponsored in part by the Naval Sea Systems Command and the Office of Naval Research under Contract N60921-83-G-A165 through the Systems Research Center at VPI&SU.

REFERENCES

Annino, J.S. and Russell, E.C. (1979). The ten most frequent causes of simulation analysis failure - and how to avoid them!. *Simulation 32,* 6 (June), 137-140.

Balci, O. (1986a). Requirements for model development environments. *Computers & Operations Research 13,* 1 (Jan.-Feb.), 53-67.

Balci, O. (1986b). Guidelines for successful simulation studies. Technical Report TR-85-2, Department of Computer Science, Virginia Tech, Blacksburg, Va., May

Balci, O. and Nance, R.E. (1985). Formulated problem verification as an explicit requirement of model credibility. *Simulation 45,* 2 (Aug.), 76-86.

Balci, O. and Sargent, R.G. (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. *Communications of the ACM 24,* 4 (Apr.), 190-197.

Balci, O. and Sargent, R.G. (1984). A bibliography on the credibility assessment and validation of simulation and mathematical models. *Simuletter 15,* 3 (July), 15-27.

Boehm, B.W., Brown, J.R. and Lipow, M. (1976). Quantitative evaluation of software quality. In *Proceedings of the 2nd International Conference on Software Engineering* (San Francisco, Calif.), pp. 592-605.

Elmaghraby, S.E. (1968). The role of modeling in IE design. *Industrial Engineering 19,* 6 (June), 292-305.

Law, A.M. (1983). Feature article: statistical analysis of simulation output data. *Operations Research 31,* 6 (Nov.-Dec.), 983-1029.

Monahan, J.F. (1977). The accuracy of stochastic algorithms. Tech. Rep. 23502, Brookhaven National Laboratory, Upton, N.Y.

Nance, R.E. (1981). Model representation in discrete event simulation: the conical methodology. Technical Report CS81003-R, Department of Computer Science, Virginia Tech, Blacksburg, Va., Mar.

Nance, R.E. (1983). A tutorial view of simulation model development. In *Proceedings of the 1983 Winter Simulation Conference* (Arlington, Va., Dec. 12-14). IEEE, New Jersey, pp. 325-331.

Nance, R.E. and Balci, O. (1986). The objectives and requirements of model management. In *Systems and Control Encyclopedia: Theory, Technology, and Applications.* (M. Singh, Ed.) Pergamon Press, Oxford. In press.

Nance, R.E., Mezaache, A.L. and Overstreet, C.M. (1981). Simulation model management: resolving the technological gaps. In *Proceedings of the 1981 Winter Simulation Conference* (Atlanta, Ga., Dec. 9-11). IEEE, New Jersey, pp. 173-180.

Roth, P.F., Gass, S.I. and Lemoine, A.J. (1978). Some considerations for improving federal modeling. In *Proceedings of the 1978 Winter Simulation Conference* (Miami Beach, Fla., Dec. 4-6). IEEE, New Jersey, pp. 213-217.

Schlesinger, S., et al. (1979). Terminology for model credibility. *Simulation 32,* 3 (Mar.), 103-104.

Schmeiser, B. (1981). Random variate generation. In *Proceedings of the 1981 Winter Simulation Conference* (Atlanta, Ga., Dec. 9-11). IEEE, New Jersey, pp. 227-242.

Wilson, J.R. and Pritsker, A.A.B. (1978). A survey of research on the simulation startup problem. *Simulation 31,* 2 (Aug.), 55-58.

Zeigler, B.P. (1976). *Theory of Modelling and Simulation.* John Wiley & Sons, New York.

Zeigler, B.P. (1984). *Multifacetted Modelling and Discrete Event Simulation.* Academic Press, London, England.

AUTHOR'S BIOGRAPHY

OSMAN BALCI is an assistant professor of Computer Science at Virginia Polytechnic Institute and State University. He received B.S. and M.S. degrees from Boğaziçi University (Istanbul, Turkey) in 1975 and 1977, and M.S. and Ph.D. degrees from Syracuse University (N.Y.) in 1978 and 1981. He is currently the vice-chairman of ACM SIGSIM, the simulation and modeling category editor of *Computing Reviews*, and the program chairman of the conference on Simulation Methodology and Validation. His current research interests are in the areas of simulation model development environments, credibility assessment of simulation results, performance evaluation, and software engineering. Dr. Balci is a member of Alpha Pi Mu, ACM, IEEE CS, ORSA, and SCS.

Osman Balci
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-4841