Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

SOME SUBJECTIVE VALIDATION METHODS USING GRAPHICAL DISPLAYS OF DATA

Robert G. Sargent

Simulation Research Group
College of Engineering and Computer Science
439 Link Hall
Syracuse University
Syracuse, New York 13244, U.S.A.

ABSTRACT

Subjective methods for operational validity are presented that use graphical displays of histograms, box plots, and behavior graphs. These methods allow the data to be correlated, have any statistical distribution, and be limited in the number of observations. Model data are used for the reference distribution (instead of a theoretical distribution such as the t or F) and for reference to compare the system data against. These methods are very general and can be used in validating different types of models.

1 INTRODUCTION

Determining that a model is valid for its intended purpose is critically important in developing a model and doing this is called the validation process. (See, e.g., Sargent (1996) for a discussion of the validation process). One step of the validation process is performing operational validity. Operational validity (Sargent 1994) is defined as "determining that a model's output behavior has sufficient accuracy for the model's intended purpose over the domain of the model's intended applicability." Various methods and techniques are used in conducting operational validity and they are applied either subjectively or objectively.

The major attribute determining which methods and techniques to use in performing operational validity is the observability of the problem entity being modeled. (A problem entity is some system (real or proposed), idea, situation, policy or phenomena.) A problem entity is (i) observable if it is possible to collect a reasonable amount of data on its operational behavior, (ii) partially observable if the amount of data that can be collected on its operational behavior is limited, and (iii) unobservable if no data can be collected on its operational behavior. The reason that observability is so important is because it is usually necessary to compare the model and problem entity behaviors (outputs) for different experimental conditions from the model's domain of intended application in order to establish a high degree of confidence that a model is valid.

It is preferable to make the comparisons of the behavior data between the problem entity and the model in objective ways, which generally is through the use of statistical tests. However, it is often not possible to use statistical tests because (i) the assumptions of the tests cannot (or only with extreme difficulty) be satisfied, (ii) the number of observations from the problem entity is insufficient, or (iii) the behavior of the problem entity is changing, e.g, is highly nonstationary. In these situations, other approaches for comparing the data are desirable.

We present some subjective approaches for performing operational validity that uses graphical displays for comparison of system and problem entity behavior data. We believe that the use of graphical displays is the most effective way of presenting the data for subjective comparison. (Furthermore, we also believe that the use of tables for comparing data is highly ineffective.) While the approaches we present are applicable to different types of models and problem entities, we orient the presentation of our approaches and the discussions about them towards the validation of stochastic discrete event simulation models. (This assumes that the problem entities are also stochastic.)

The remainder of this paper is organized as follows: Section 2 contains some different ways of displaying data graphically, Section 3 discusses some approaches for performing operational validity that use graphical displays of data, and Section 4 gives the summary.

2 GRAPHICAL DISPLAYS

In this section we present three different ways of displaying data graphically: histograms, box plots, and behavior graphs. These three types of displays (or graphs) allow data to be statistically dependent (i.e.,

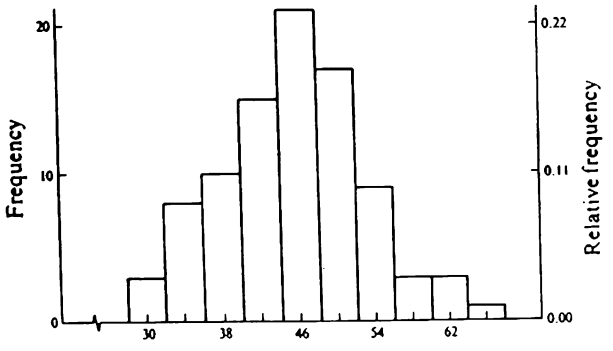

Figure 1: Histogram

correlated), which often occurs in behavior data of systems and simulation models.

2.1 Histograms

Histograms are used to display data for frequency and relative frequency distributions. To construct a histogram, the data are grouped into nonoverlapping intervals (or classes) of usually equal width. The number of intervals is usually between seven and fifteen, depending on the total number of observations. The area of each interval (class) is proportional to the number of observations in each interval. If the interval widths are equal, the height of each interval of a histogram of a frequency distribution is equal to the number of observations in the interval, and the height of each interval of a histogram of a relative frequency distribution is equal to the number of observations in the interval divided by the total number of observations. See Figure 1 for an example of a histogram.

The data plotted in a histogram should be identically distributed. (Note: Prior to putting data in a histogram, the data should be plotted in a time-sequence to ensure that there is no time trend in the data (Johnson 1994), which would cause the data to be not identically distributed.) The data may be statistically independent or dependent (i.e., correlated) (Box, Hunter, and Hunter 1978). The data plotted may be the observations collected on the system (or from the simulation model) or may be some function of partitioned subsets of the observations such as sample means.

2.2 Box Plots

A box plot (Johnson 1994) or box and whisker plot (Walpole and Myers 1993) depict the three quartiles and the two extreme values of a set of data. A rectangle box encloses the 25th percentile (lower quartile)

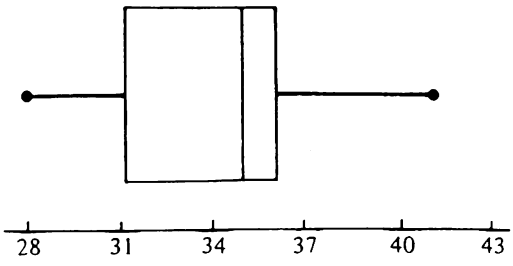

Figure 2: Box Plot

and the 75th percentile (upper quartile) and a line sectioning the box displays the 50th percentile (median). A whisker at each end of the box extends to an extreme value. See Figure 2 for an example.

The data used for box plots should be identically distributed. The data may be statistically independent or dependent (i.e., correlated.). The data used may be the observations themselves or some function of partitioned subsets of the data such as sample means.

2.3 Behavior Graphs

A behavior graph shows the relationship between two entities such as parameters, variables, and functions of random variables by plotting paired data on the two entities. The data points plotted may be the observations themselves, which may be few in number, or may be functions of partitioned subsets of the observations, which may be based on a large number of observations. Some examples of different types of relationships that may be shown in behavior graphs are relationships between two deterministic variables, between two functions of a single random variable such as its mean and standard deviation, between a parameter and some function of a random variable, and between a function of one random variable and a function of another random variable.

Figures 3 and 4 contain two examples of behavior graphs. They are from Anderson (1974) and are based on observations collected on an interactive computer system. Figure 3 contains data showing the relationship between the mean and the standard deviation of system reaction time. One can readily see that there is a linear relationship between them. Each data point is based on thirty independent observations. Figure 4 shows the relationship between the average length of a time slice and the total number of disk accesses based on twenty (correlated) observations per data point. A linear relationship between these two entities can be seen in the behavior graph.

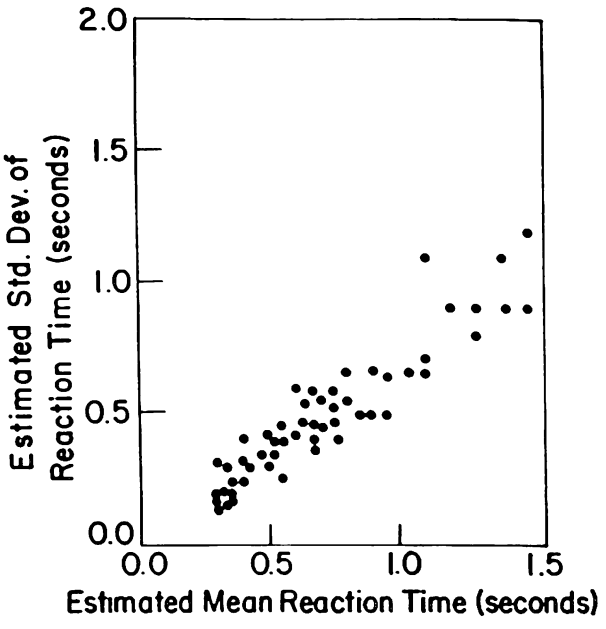

Figure 3: Mean vs. Standard Deviation

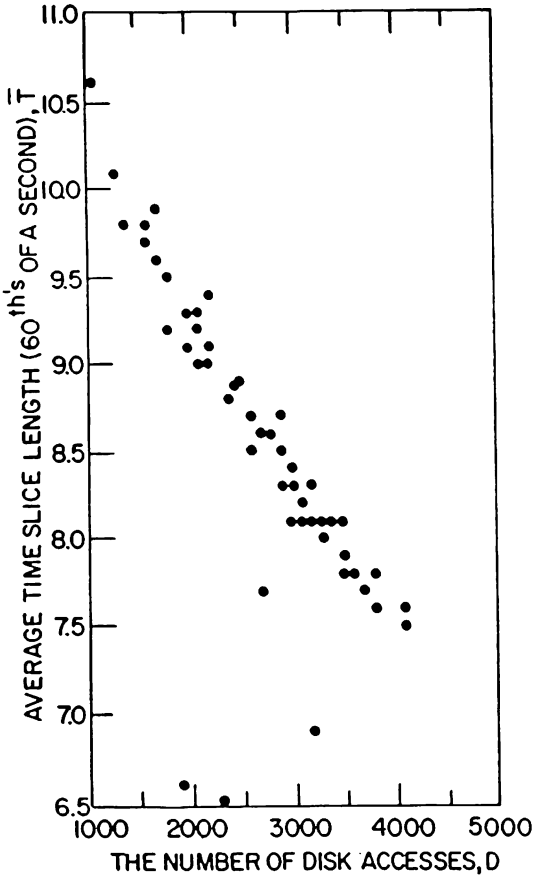

Figure 4: Average Time vs. Disk Accesses

3 VALIDATION APPROACHES

In operational validity, comparisons are made whenever possible between data from the problem entity and the model to determine if the model has sufficient accuracy for the model's intended purpose over the model's domain of intended applicability. It is preferable to do this in some objective way, which means using some type of statistical test if the system (or simulation model) is stochastic. However, this is often not possible because (i) the assumptions required of the statistical tests cannot be satisfied (or only with great difficulty) and/or (ii) there are insufficient data available from the system. Most (simple) statistical tests require the data to be independent, which is often not true of data collected from a system or simulation model. In addition, many statistical tests require the data to have a normal distribution, which is also usually not the case for data collected from a system or simulation model. Furthermore, even when the data does satisfy the statistical requirements, statistical tests frequently cannot be used for operational validity because there is insufficient number of observations from the system to obtain "meaningful" results from the statistical test. For example, the length of a confidence interval developed to compare the means of some model and system behavior variable can be to long for any practical usefulness. This can be caused by an insufficient number of system observations. Thus, it is often the case that statistical tests cannot be used for operational validity. We present in this section some approaches that can be used for operational validity, which eliminates the requirements of independence in the data, have no distributional requirements on the data, and can be used with limited number of system observations.

Statistical tests use a theoretical reference distribution such as the t or F distribution (Box, Hunter, and Hunter 1978; Johnson 1994; Walpole and Myers 1993). In the approaches presented for operational validity in this paper, data generated from the model are used for the reference distribution (or reference) instead of a theoretical distribution. (See Box, Hunter, and Hunter (1978) for a discussion on using data as a (external) reference distribution instead of a theoretical reference distribution.) The data generated from the model for use as a reference distribution (or reference) is displayed in one of the graphical displays described in Section 2 along with the data from the system. These two sets of data are compared subjectively to determine whether the model has sufficient accuracy for its intended purpose. This comparison can be make by the model development team and/or by experts using face validity or Turing tests.

In making comparisons between the model and system data to decide model validity, two types of incorrect decisions can be made. One is to reject a valid model as being invalid and the other is to accept an invalid model as being valid. The first type of incorrect decision is a type I error and is the model builder's risk, and the second type of incorrect decision is a type II error and is the model user's risk (Balci and Sargent 1981). The type II error is extremely important and should be minimized.

3.1 Histograms as Reference Distributions

Histograms of data generated from a model can be used as reference distributions for making comparisons between model and system data for operational validity. For each entity of interest (e.g., an output random variable or sample mean of some output random variable), a sufficient number of observations is generated from the model to have at least fifty data points to be placed in a histogram. The model data to be placed in each histogram need only be identically distributed. Similarly, for each entity of interest, system observations are used to obtain system data points. The system data points must also be identically distributed for each entity of interest. The number of system data points for each entity of interest is usually just a few and may be only one. For each entity of interest, a histogram of the model data, for use as the reference distribution, and a histogram of the system data are placed in the same figure to be compared subjectively to aid in deciding whether the model's entity of interest has sufficient accuracy for the model's intended purpose.

As an example, consider a simulation project by Lowery (1996). A simulation model was developed to predict the mean (average) number of beds used daily (census) in specific hospital units. Operational validity was performed to determine whether the simulation model's mean census (usage) of beds were within the required accuracy of four beds for large hospital units (i.e., units having a large number of beds). Since there was a day a week effect, only Mondays will be considered here. There were 24 system observations (weeks) available on Monday census for the unit that we consider and these observations are correlated. The data entities of interest to be compared must be determined. We select two for illustration purposes: a 24-week average daily (Monday) census and a 4-week average daily census. Observations were generated from the simulation model to obtain fifty 24-week average daily census (for Mondays) data points. A histogram of the 24-week average daily census is in Figure 5. This histogram is the reference distribution for the one system data point of a 24-week average daily census. One can readily see that this system data point lies within the reference distribution. Figure 6 contains a histogram of fifty 4-week average daily census (for Mondays) data points developed from observations from the simulation model for use as a reference distribution. The system observations are used to create six 4-week average daily census and a histogram of them is also placed in Figure 6. One can readily observe that the system histogram lies within the reference distribution. Thus, based on these two figures the simulation model can be judged to have sufficient accuracy with respect to mean daily census for this hospital unit on Mondays. Note that the only assumptions required for this approach are (i) the data compared must be on the same entity of interest, (ii) the data from the model have to be identically distributed, and (iii) the data from the system have to be identically distributed.

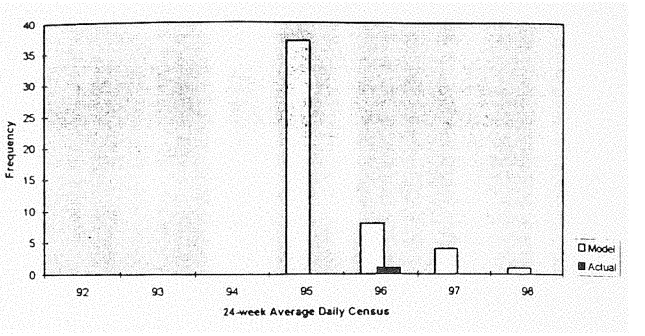

Figure 5: 24-weekly Average Daily Census

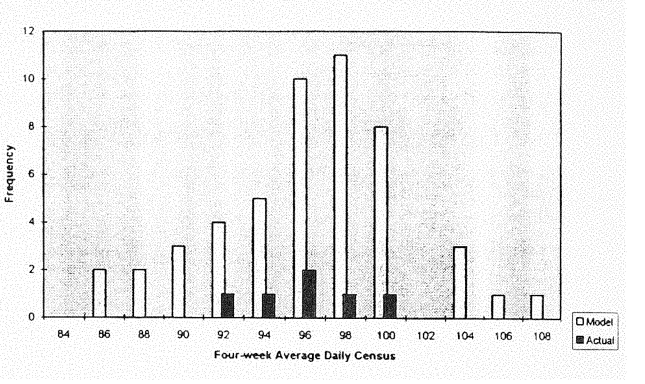

Figure 6: 4-week Average Daily Census

3.2 Box Plots as References

Box plots of data generated from a model can be used as references for making comparisons between model

and system data for operational validity. For each entity of interest (e.g., an output random variable or sample mean of some output random variable), a sufficient number of observations is generated from the model to have at least fifty data points to develop a box plot. The model data points used to develop the box plot need only to be identically distributed. Similarly, for each entity of interest, system observations are used to develop system data points, which must be identically distributed. The number of system data points should be at least ten (hopefully) and preferably thirty or more. For each entity of interest, a box plot of the model data to be used as a reference, and a box plot of the system data are placed in the same figure to be compared subjectively. (To assist in making comparisons of means using box plots, a rule of thumb given in Walpole and Myers (1993) may be helpful: "a rough guideline is that if the 25th percentile line for one sample exceeds the median line for the other sample, there is strong evidence of a difference between the means." Note that a comparison can show that a model is invalid but cannot *prove* that a model is valid.)

An example of a box plot used in performing operational validity is given in Figure 7. The box plots are Sunday census observations for the same hospital unit discussed above. The model box plot, which is the reference, is developed from fifty observations (Sunday census) generated by the simulation model. The system box plot is developed from the 24 observations (Sunday census) collected on the hospital unit. In comparing the two box plots, it appears that the model has more variability in its Sunday census than the hospital unit. Regarding the mean census, it is this author's opinion that this pair of box plots shows insufficient evidence to judge that the model's mean census is not within four beds of the hospital's mean census; i.e., there is insufficient evidence to reject the model as being invalid. (In performing operational validity on this model, several different comparisons were made. This is an example of one of them.)

We note that box plots only require identically distributed data. Box plots are extremely effective in communication and are thus effective for conveying information on model validation and for helping with model acceptability (Sargent 1996). For example, a pair of box plots for each day of the week for the hospital unit discussed above could be put into a single graph to be used for communication to users regarding the validity of this simulation model. The use of box plots for operational validity may require more system data than the use of histograms, which were discussed above.

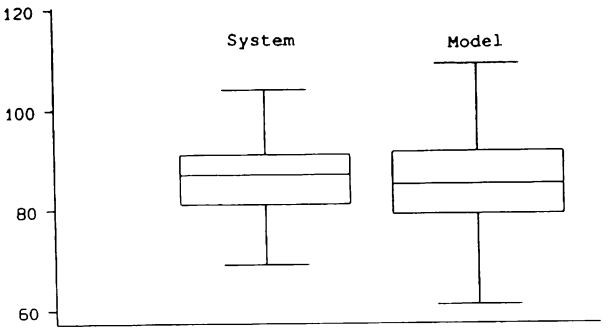

Figure 7: Box Plots of Sunday Census

3.3 Behavior Graphs as References

In performing operational validity, comparisons of different behavior relationships occuring in the system should be made to those occuring in a model. We suggest the use of behavior graphs as one approach to doing this. It is often difficult to use objective validation methods such as statistical tests because the system/model behavior may be nonstationary, may operate over a large portion of the application domain, and data may be correlated. Behavior graphs avoid the use of statistical tests by using subjective analysis and model data as references.

Behavior graphs can be used to show different types of relationships as discussed in Subsection 2.3. Different types of relationships require different amounts of system data. If operational validity is being performed on a deterministic model of a deterministic system, deterministic relationships are used and these generally require only a few observations. In performing operational validity on a stochastic model of a stochastic system, a large number of system observations are often needed. For example, stochastic simulation models of computer and communication systems generally have model and system relationships developed from a large number of observations from the model and from the system.

To illustrate behavior graphs, we consider a simulation model of an interactive computer system in Anderson and Sargent (1974) where behavior graphs were used to validate the model. Three behavior graphs they used are presented in Figures 8, 9 and 10. The relationship between the mean and standard deviation of reaction time is shown in Figure 8. One can readily observed that the same linear relationship occurs in both the model and the system. Figure 9 contain relationships for both the maximum observed value and the average value of reaction time versus the total number of disk accesses. Each data point

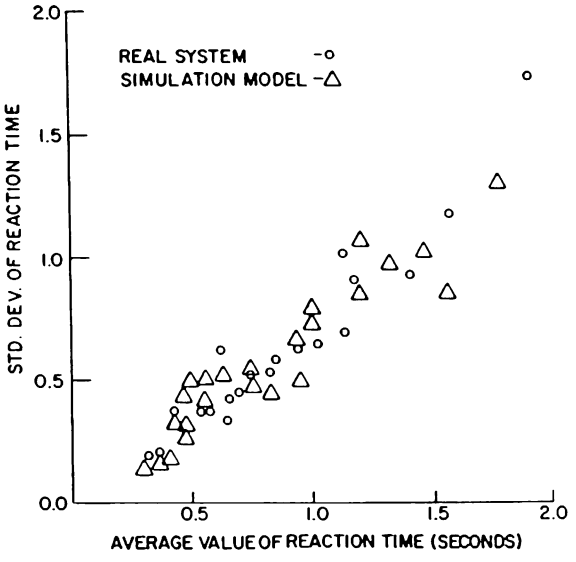

Figure 8: Mean vs. Standard Deviation

is from (or represents) five minutes of computer system time. We observe that these model and system relationships are similar with the exception that the system has more variability than the model. Figure 10 contains the relationship of average response time versus average background queue length. One can readily observe that these model and system relationships are similar except for two system data points, which is important to ask why". (For details on these graphs and the validation of this simulation model, see Anderson (1974) and Anderson and Sargent (1974).)

4 SUMMARY

Three different types of graphical displays were presented that have minimal assumptions required of the data. Methods for operational validity that use these graphical displays were described. An important feature of these methods is that model data is used for the reference distribution (or reference) instead of a theoretical (statistical) distribution for the system data to be compared against. The graphical methods presented should provide significant help in performing operational validity since the use of graphs is the most used approach in performing operational validity (Sargent 1996).

ACKNOWLEDGMENT

This author acknowledges Julie Lowery for providing the figures from the hospital simulation study used in this paper.

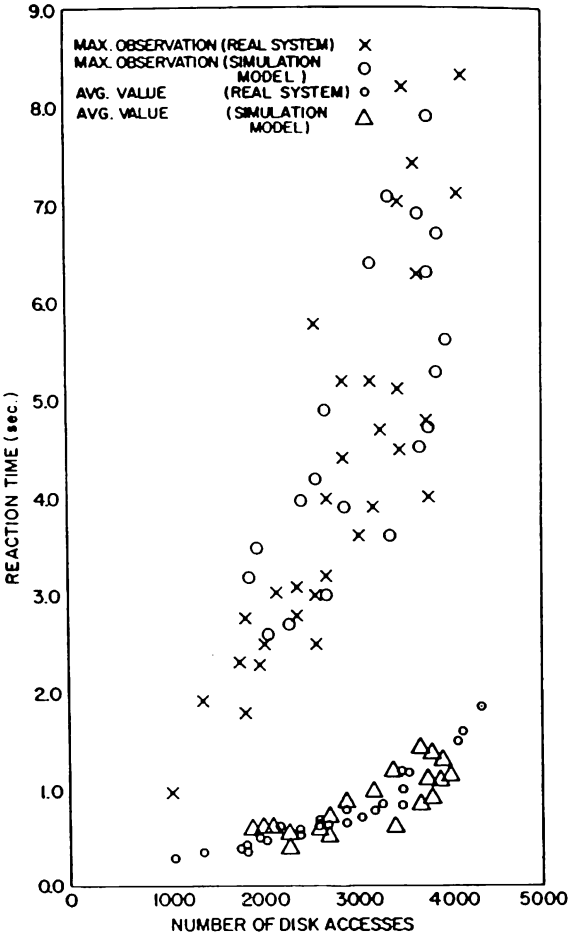

Figure 9: Reaction Time vs. Disk Accesses

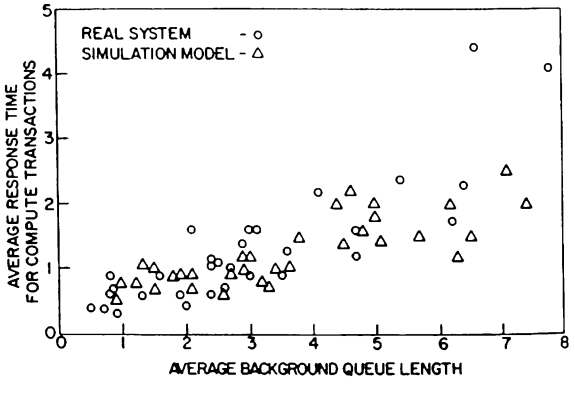

Figure 10: Response Time vs. Queue Length

REFERENCES

Anderson, H. (1974). *An empirical investigation into foreground-background scheduling for an interactive computer system.* Ph. D. thesis, Syracuse University.

Anderson, H. and R. Sargent (1974). An investigation into scheduling for an interactive computer system. *IBM journal of research and development 18*(2), 125–137.

Balci, O. and R. Sargent (1981). A methodology for cost-risk analysis in the statistical validation of simulation models. *Comm. of the ACM 24*(4), 62–71.

Box, G., W. Hunter, and J. Hunter (1978). *Statistics for experimenters.* John Wiley and Sons.

Johnson, R. (1994). *Miller and Freund's probability and statistics for engineers* (5th ed.). Prentice Hall.

Lowery, J. (1996). Design of hospital admissions scheduling system using simulation. In *Proceedings of the 1996 Winter Simulation Conference*, J.M. Charnes, D.J. Morrice, D.T. Brunner, and J.J. Swain (eds.).

Sargent, R. (1994). Verification and validation of simulation models. In *Proceedings of the 1994 Winter Simulation Conference*, J.D. Tew, S. Manivannan, D.A. Sadowski, and A.F. Seila (eds.), Lake Buena Vista, FL, pp. 77–87.

Sargent, R. (1996). Verifying and validating simulation models. In *Proceedings of the 1996 Winter Simulation Conference*, J.M. Charnes, D.J. Morrice, D.T. Brunner, and J.J. Swain (eds.).

Walpole, R. and R. Myers (1993). *Probability and statistics for engineers and scientists* (5th ed.). Macmillan Publishing Company.

AUTHOR BIOGRAPHY

ROBERT G. SARGENT is a Professor at Syracuse University. He received his education at the University of Michigan and has published widely. Dr. Sargent has served his profession in numerous ways and has been awarded the TIMS College on Simulation Distinguished Service Award for long-standing exceptional service to the simulation community. His research interests include the methodology areas of modeling and discrete event simulation, model validation, and performance evaluation. Professor Sargent is listed in *Who's Who in America.*