COMMITTEE T1
CONTRIBUTION

Document Number T1A1.5/92-137

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: Confidence Intervals for Correlation Coefficients

****

ISSUE ADDRESSED: Video Quality Tests

****

SOURCE: NTIA/ITS - Arthur Webster

****

DATE: July 13, 1992

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS: Video Teleconferencing, Video Telephony, Video Quality, Subjective Quality, Objective Quality

****

DISCLAIMER:

****

Confidence Intervals for Correlation Coefficients

1. Introduction

This contribution presents the results of a study done to evaluate the degree of confidence we can have in the correlation coefficients calculated between subjective and objective video quality scores (see for example References [1] and [2]). This contribution provides a clear, objective way to determine the number of systems that should be included in a performance test. We recommend that this contribution be inserted into the Appendix of the Draft Standard [3] under the heading titled "Test Results."

2. Theoretical Background

The confidence interval on the correlation coefficient, ρ, has been investigated by those interested in objective measures of speech quality (see Ref. [4]). A useful equation has been found which closely approximates the true confidence interval for a large number of data points (i.e. $N \geq 25$). See Ref. [4] and [5]. In our case, given the currently proposed test methodology (see Ref.[3]), each point corresponds to an (x,y) pair, where x is the subjective rating of the system under test and y is a proposed objective measurement. Correlation between subjective and objective scores is one method of comparing different objective measurements. It is this measure, ρ, which is addressed in this contribution. The closer $|\rho|$ is to 1.0, the better the objective measure.

The $100(1-\alpha)$ percent confidence interval (α=0.05 for a 95% confidence interval) for ρ is given in References [4] and [5] as

$$\tanh\left(arctanh(\hat{\rho}) - \frac{Z_{\alpha/2}}{\sqrt{N-3}} \right) \leq \rho \leq \tanh\left(arctanh(\hat{\rho}) + \frac{Z_{\alpha/2}}{\sqrt{N-3}} \right) \quad (1)$$

where ρ is the true correlation coefficient, $\hat{\rho}$ is the correlation coefficient estimated from the data, $Z_{\alpha/2}$ is the normal(Gaussian) deviate (see Ref.[5] p. 462-3), N is the number of points (in our case, the number of systems under test), and **tanh** is the hyperbolic tangent.

3. Monte Carlo Simulation

Since the above equation is considered valid for $N \geq 25$ and the T1A1.5 VTC/VT sub-working group might be interested in a smaller number of systems for the upcoming VTC/VT test, a computer simulation was performed to both validate Equation 1 and to extend the curves into our region of interest.

In order to determine the confidence interval for a given correlation coefficient, we began by constructing four large (1,000,000 points) pairs of number sequences with known correlation coefficients of ρ = .85, .90, .95, and .99. This was accomplished by utilizing the Unix-C random number generator **drand48** and constructing the pair of sequences according to the following:

$$
\begin{aligned}
x &= UNIFORM\ (0,1) \\
z &= UNIFORM\ (0,1) \\
y &= \alpha\, x + \beta\, z \qquad (2) \\
\rho_{xy} &= \frac{\alpha}{\sqrt{\alpha^2 + \beta^2}}
\end{aligned}
$$

The data from the simulation was collected by calculating $\hat{\rho}$ over 10,000 sequences of (x,y) pairs with each sequence containing N pairs. Each sequence was drawn from one of the large number sequences (described above in Equation 2) that have known correlation coefficients. From these 10,000 estimates of ρ, the confidence intervals are obtained by finding the upper and lower boundaries within which 90%, 95%, and 99% of the estimates are found. N was varied from 5 to 100.

The twelve cases that were considered involved combinations of four correlation coefficients (.85, .90, .95, .99) and three confidence intervals (90%, 95%, 99%).

4. Results

Figure 1 shows the case where ρ=.95 and the confidence interval is 90%. Note that the simulation curve and the theoretical curve (from Equation 1) converge as N increases. Figures 2 through 7 show the twelve cases for N between 5 and 20.

5. Ramifications on Test Design

This contribution offers an excellent method that the T1A1.5 VTC/VT sub-working group can use when designing its performance tests. It is important to note that a correlation coefficient estimate is not statistically different from any other within its confidence interval [4]. That means that to have a statistically significant result from a test a certain minimum number of systems must be used. For example, with a 90% confidence interval for a ρ of .90 and only 5 systems (see Figure 2b), the true correlation coefficient may be as low as .63. One cannot claim that this is statistically different from another result which actually did give .63. Also, as can be seen from the plots, improvement in the confidence interval diminishes beyond N=15 or so. The knee of the curve is around N=10. Therefore the number of systems chosen should be at least 10 to 15 to ensure a statistically valid result.

If we assume that all test scenes will be used for each system chosen, then the number of clips to be subjectively rated will be (Number of Systems) x (Number of Scenes). The ITS subjective test used 146 scene/system clips in four twenty minute viewing sessions. Each scene was 9 seconds long and we used 48 viewers. For a single test, a reasonable number of clips would be between 150 and 250. The CCIR Recommendation 500-3 limits the length of time for a viewing session to half an hour and the number of sessions per viewer to 4. Therefore the scene length must also be considered. If we choose 15 systems we might choose 10 to 16 test scenes.

6. Conclusions

This contribution offers an objective way to choose the number of systems needed for a statistically valid quality test (video or otherwise). We are aware of the need to choose test scenes which appropriately span the spatial-temporal information matrix (See Refs. [6], [7], and [8]). This contribution shows the importance of choosing an adequate number of systems for the test. If more than 250 scene/system combinations are desired for the test, it would be useful to consider multiple tests. That is, test half of the clips with one group of viewers and the other half with another group of viewers.

We recommend that this contribution be inserted into the Appendix of the Draft Standard under the heading titled "Test Results."

7. References

[1] T1A1.5/92-112, Voran,S. and Wolf,S., Objective Measures of Video Impairment: Analysis of 128 Scenes, March 25, 1992.

[2] T1Q1.5/92-112, Voran,S., Objective Measures of Video Impairment: An Update on the ITS Derivation Process.

[3] T1A1.5/92-107, Appendix A to Draft American National Standard, Digital Transport of Video Teleconferencing/Video Telephony Signals- System M-NTSC Analog Interface Specifications and Performance Parameters, March 4, 1992.

[4] Quackenbush,S., Barnwell,T., and Clements, M., Objective Measures of Speech Quality, Prentice Hall, 1988, p. 191.

[5] Montgomery,D. and Peck,E., Introduction to Linear Regression Analysis, John Wiley, 1982, p. 49.

[6] T1Q1.5/92-13, Webster, A. & Wolf, S., Spatial and Temporal Information Measures for Video Quality, January 22, 1992.

[7] T1A1.5/92-13, Webster, A., Spatial and Temporal Information Measures -- Test Scene Evaluation, March 31, 1992.

[8] T1A1.5/92-134, Webster, A., Spatial and Temporal Information Measures -- Test Scene Evaluation II, July 13, 1992.

Figure 1. Confidence Intervals for Correlation Coefficients

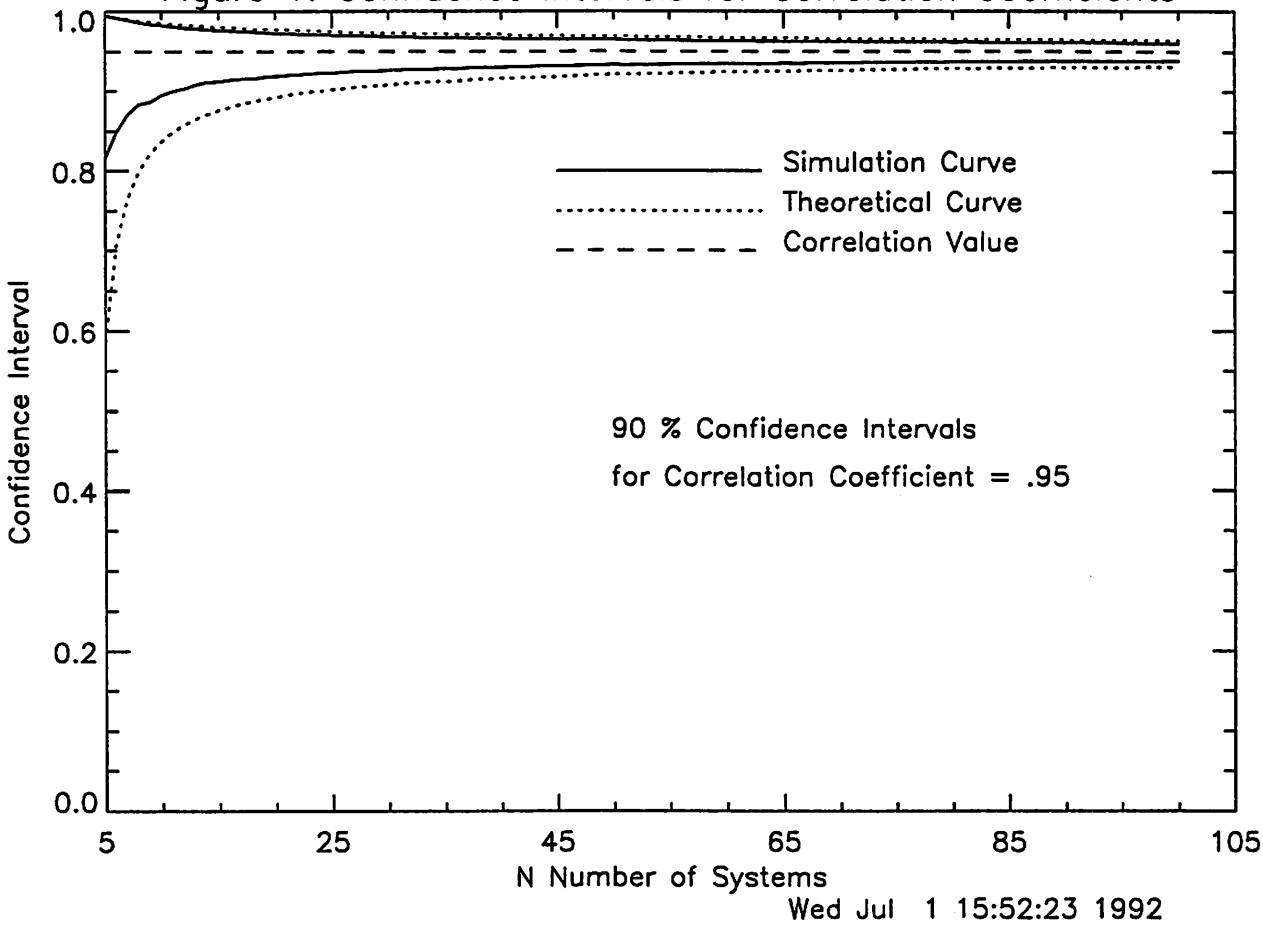

Figure 2a. Confidence Interval = 90% Correlation Coefficient = .85

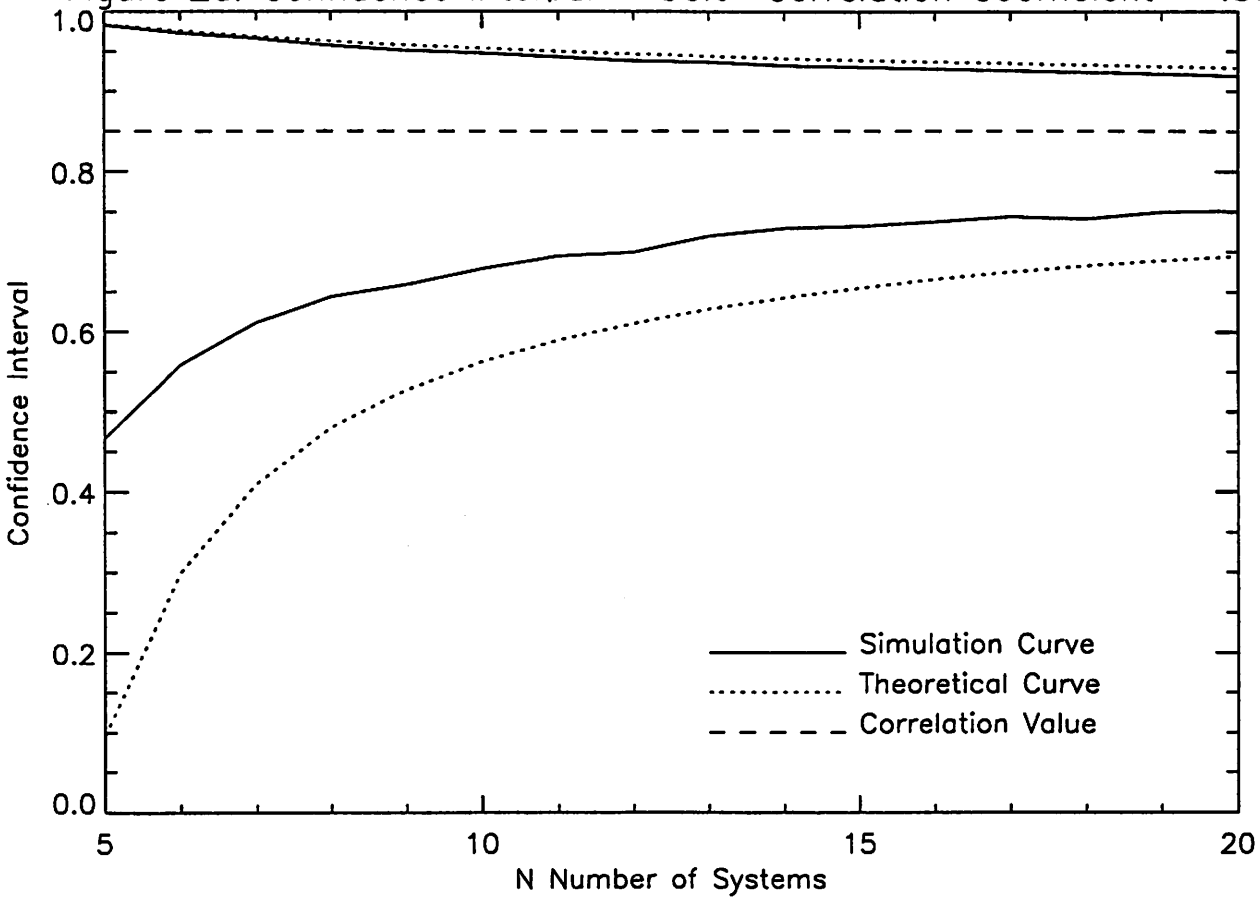

Figure 2b. Confidence Interval = 90% Correlation Coefficient = .90

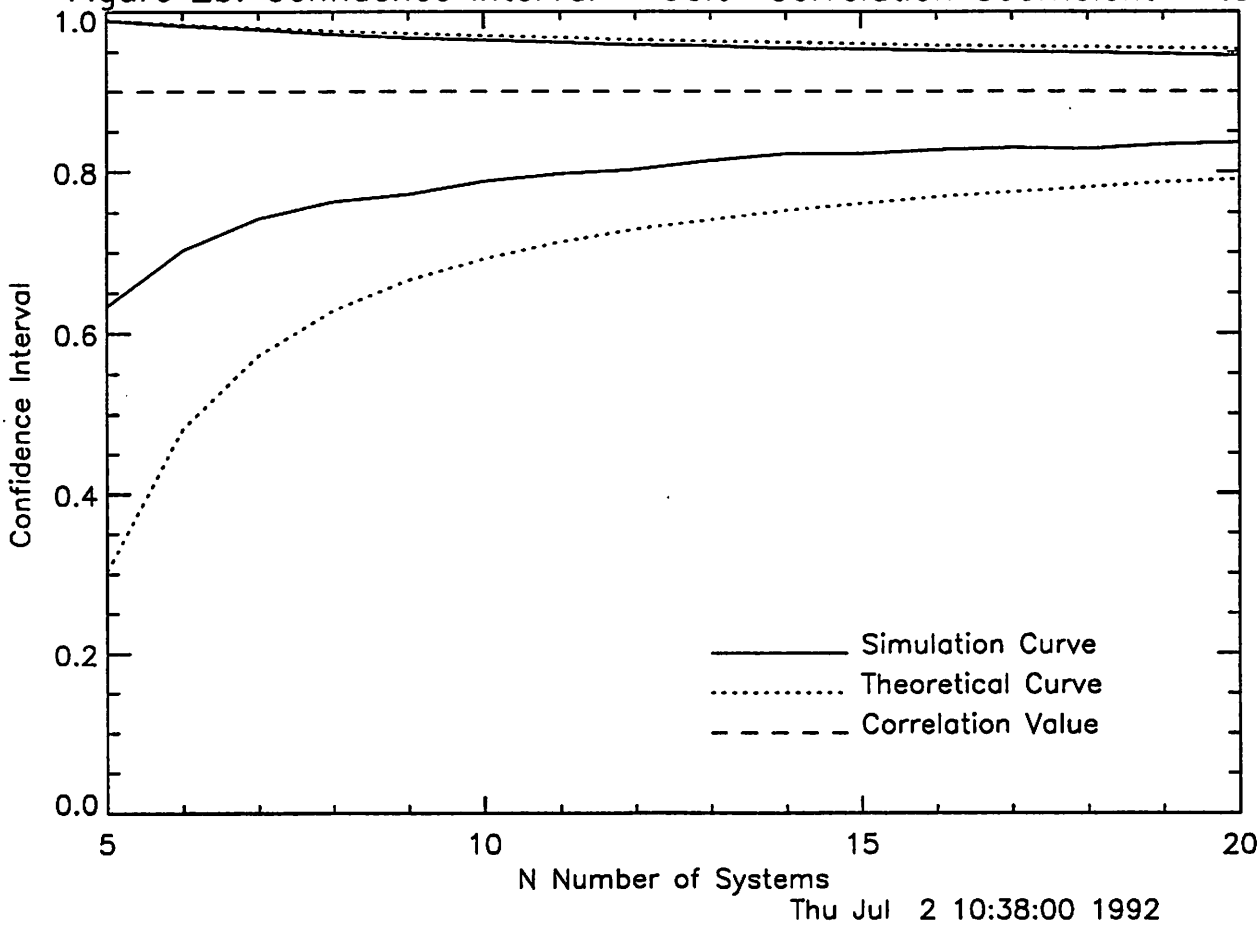

Figure 3a. Confidence Interval = 90% Correlation Coefficient = .95

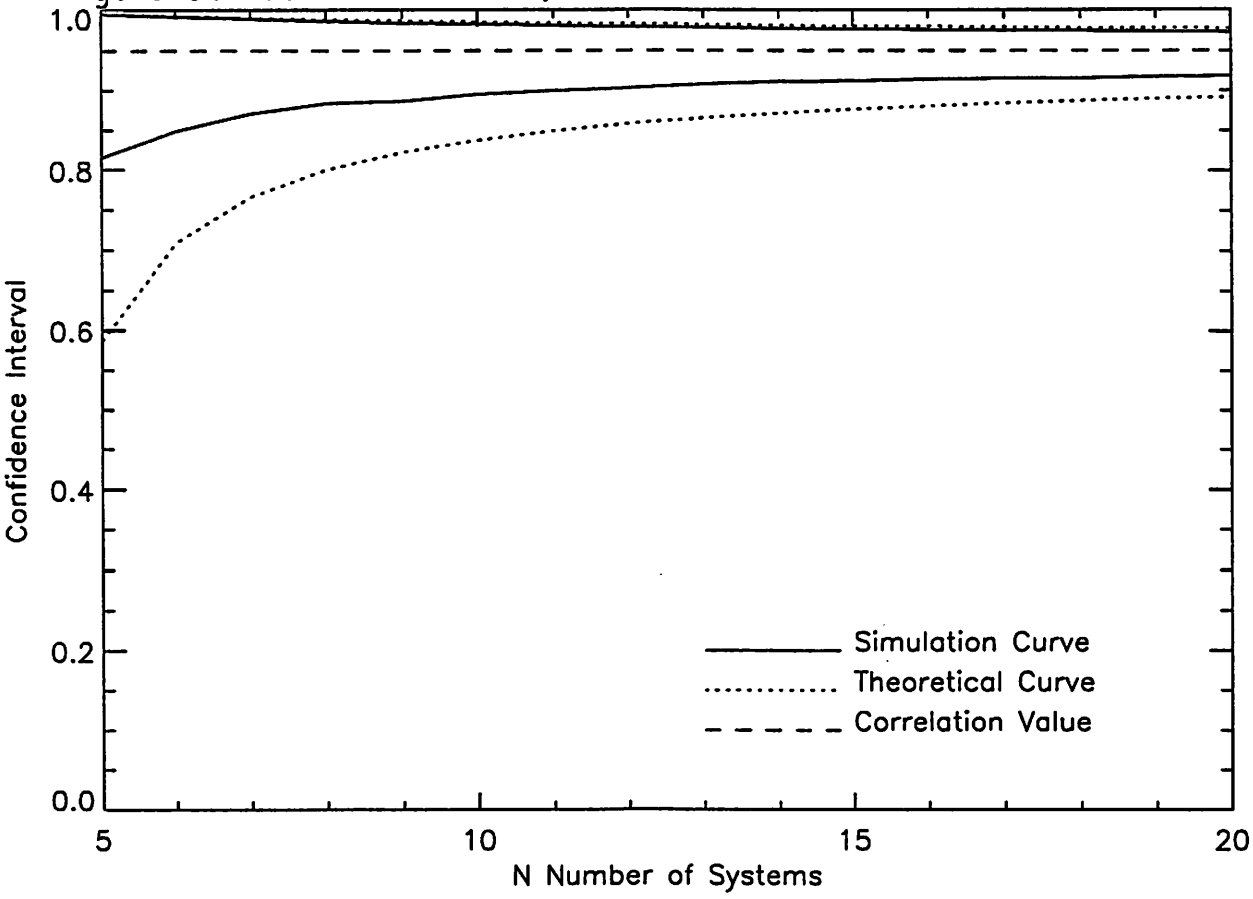

Figure 3b. Confidence Interval = 90% Correlation Coefficient = .99

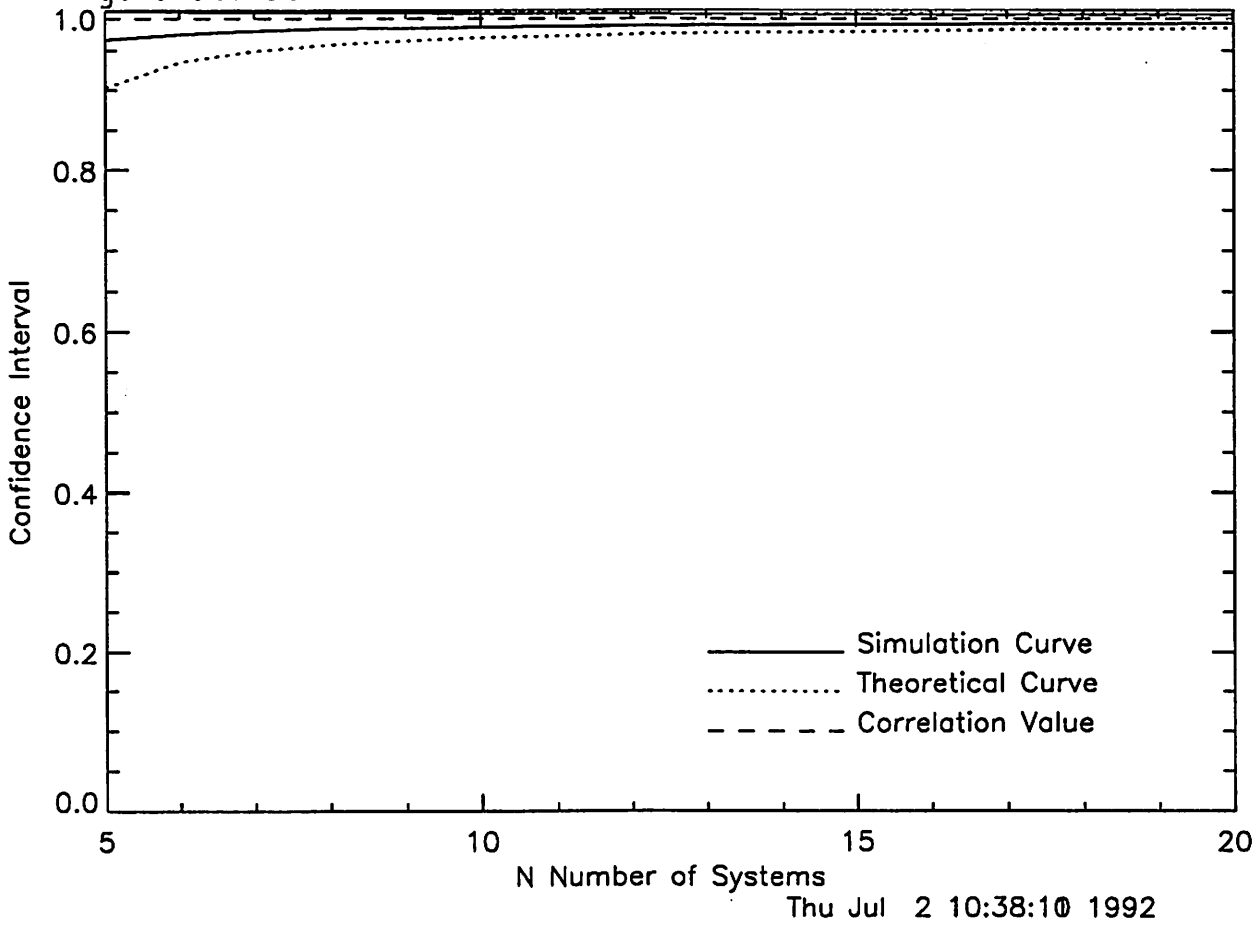

Figure 4a. Confidence Interval = 95% Correlation Coefficient = .85

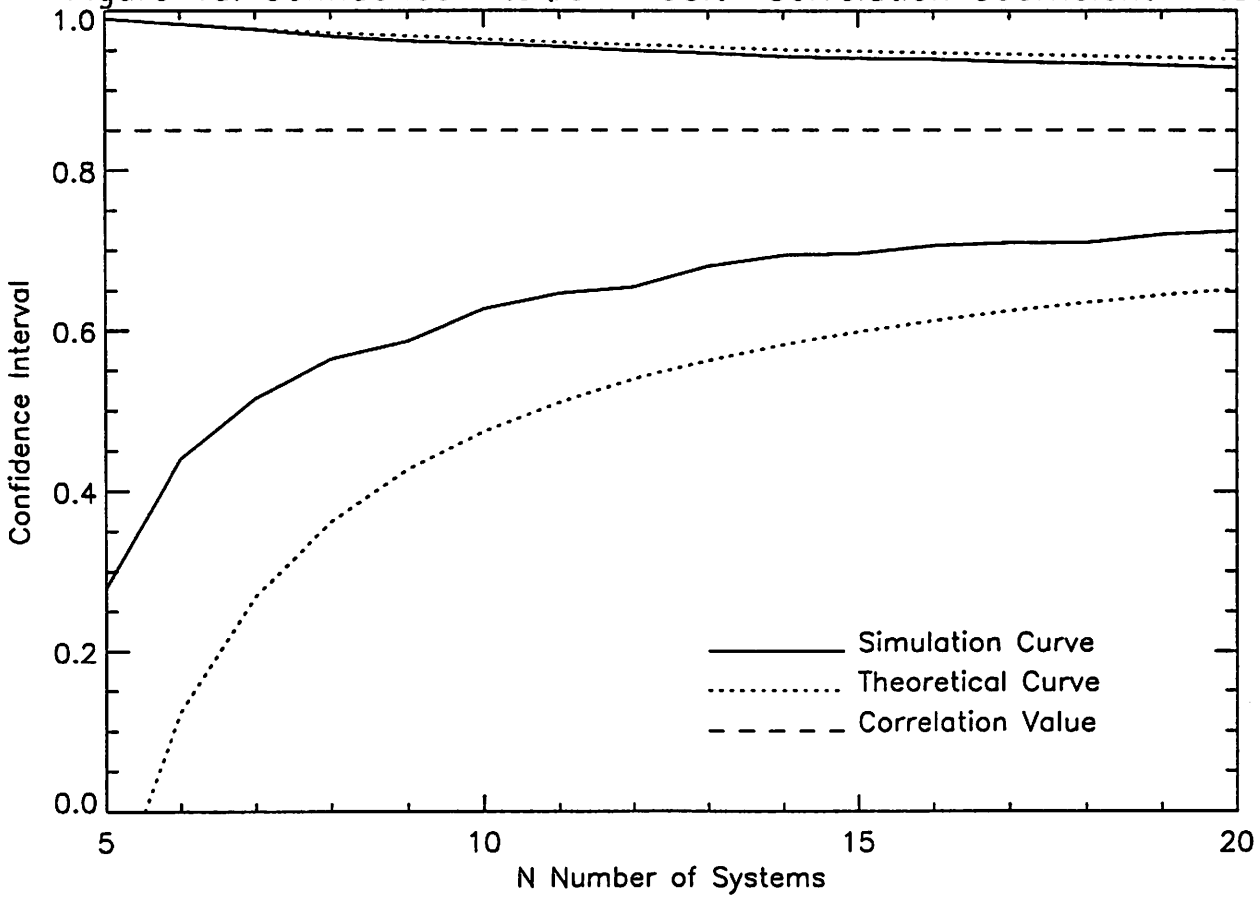

Figure 4b. Confidence Interval = 95% Correlation Coefficient = .90

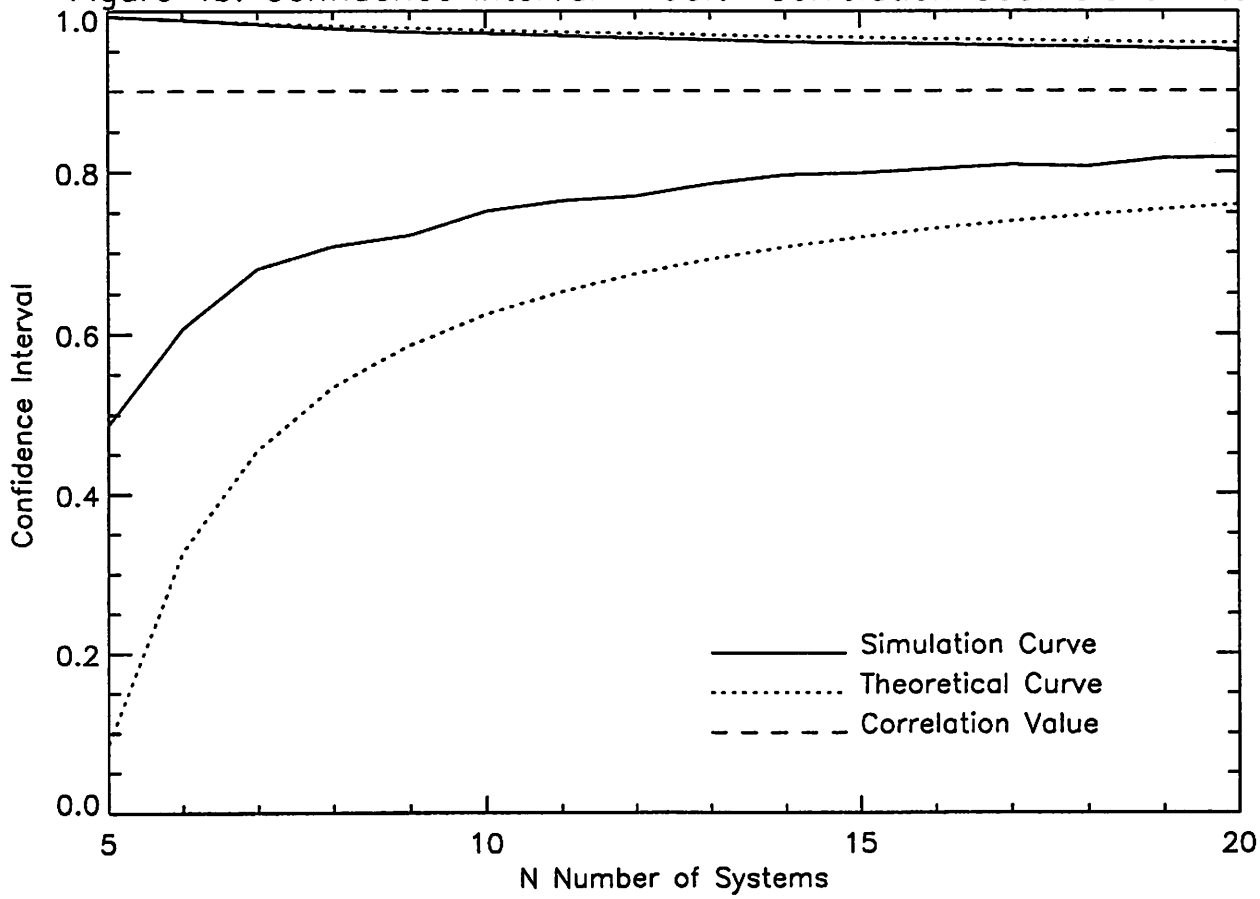

Figure 5a. Confidence Interval = 95 Correlation Coefficient = .95

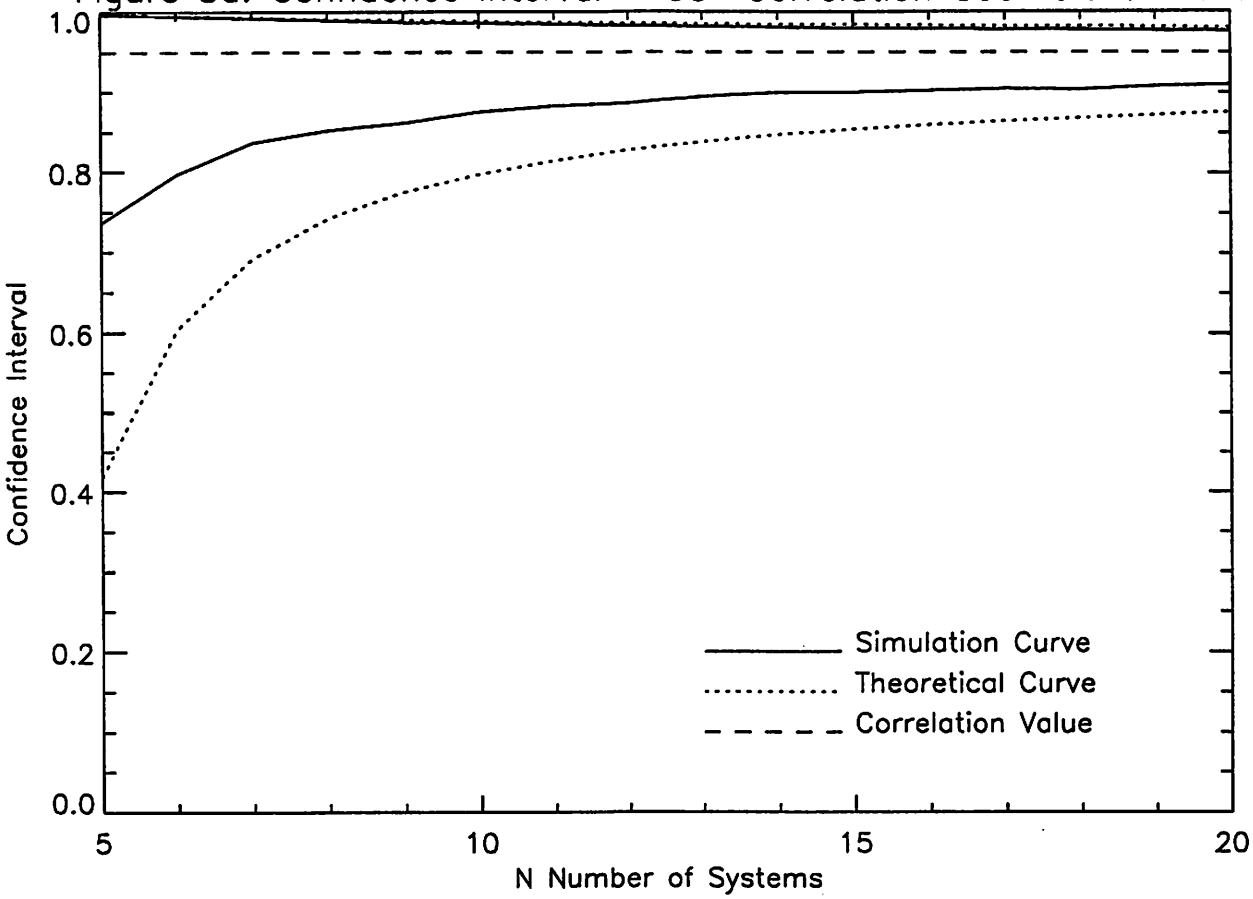

Figure 5b. Confidence Interval = 95% Correlation Coefficient = .99

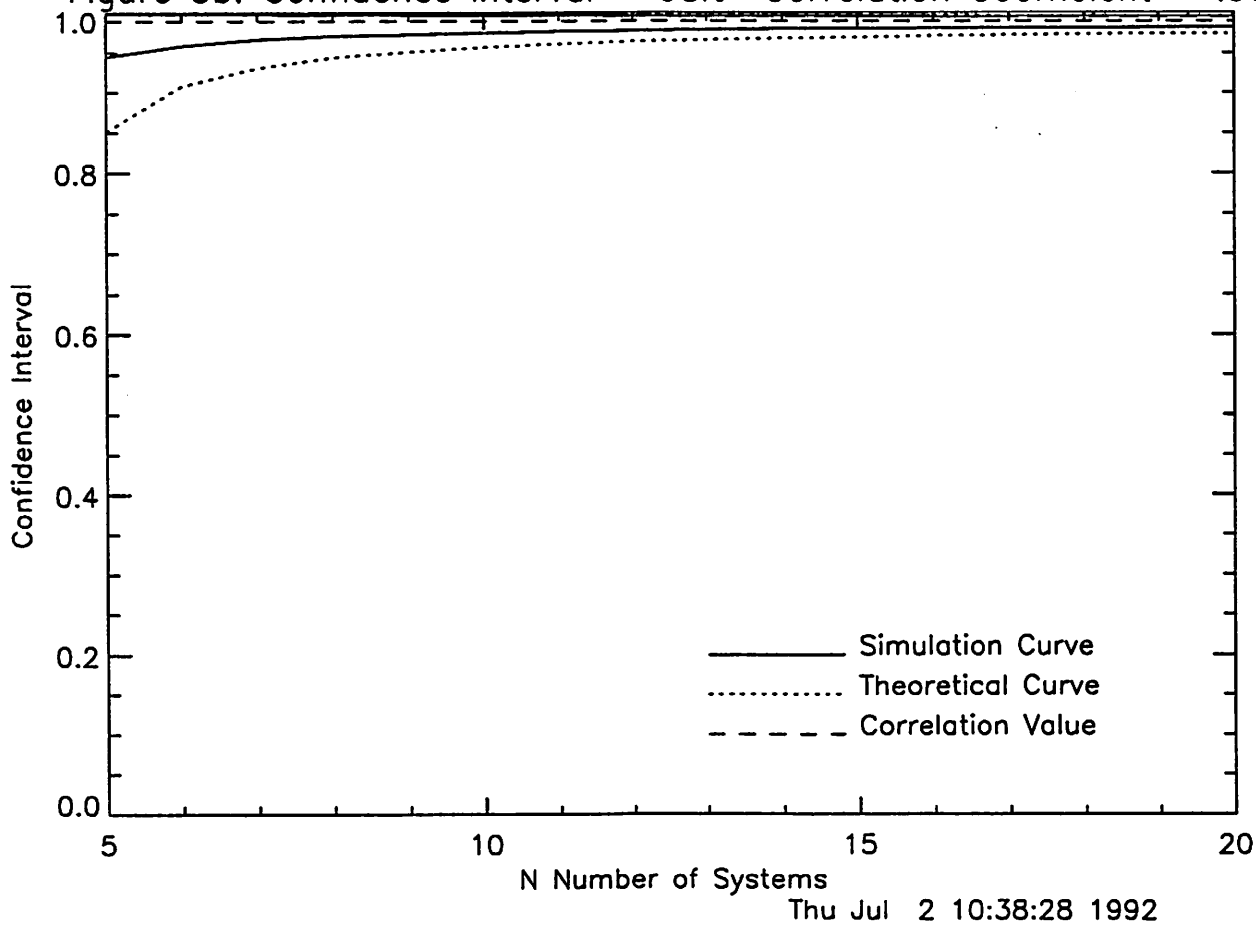

Figure 6a. Confidence Interval = 99% Correlation Coefficient = .85

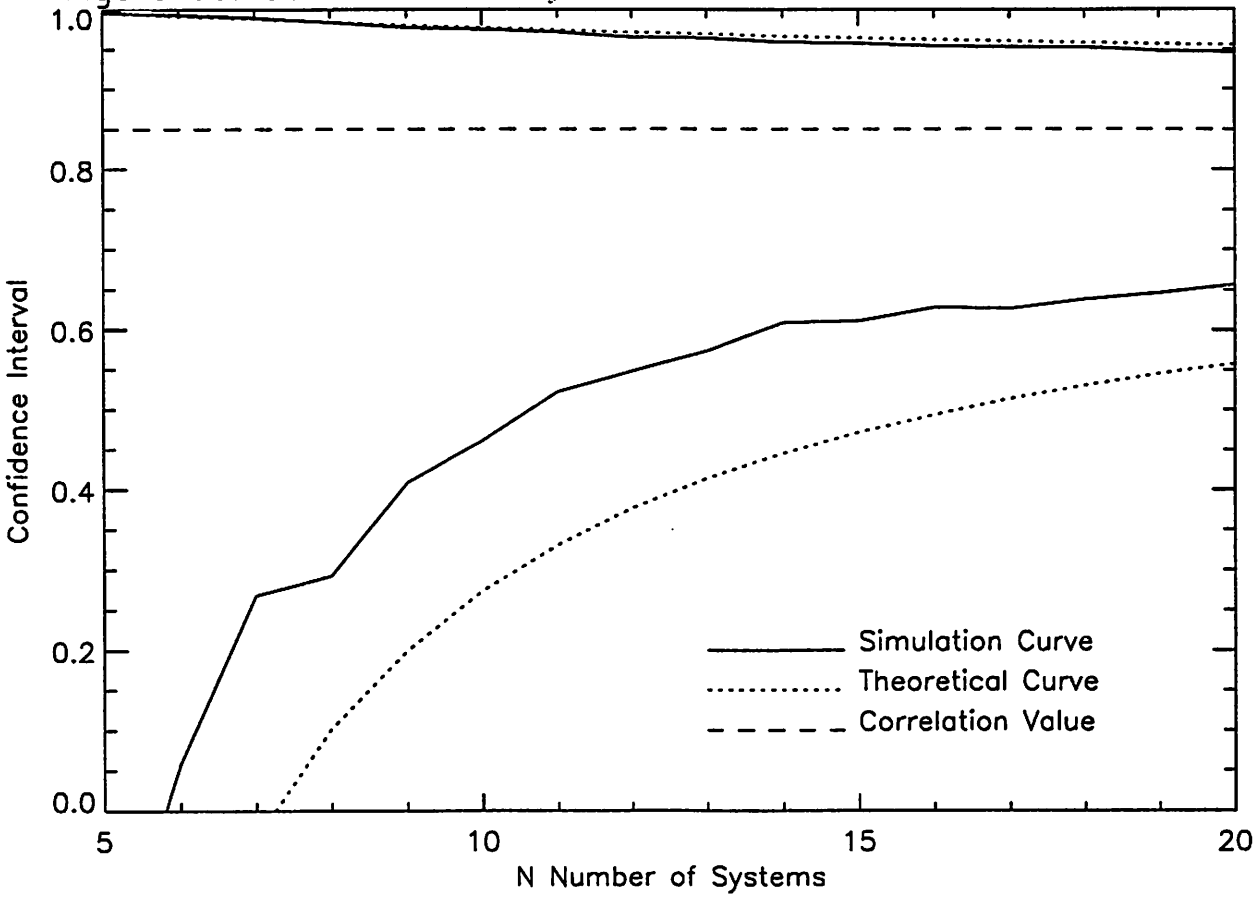

Figure 6b. Confidence Interval = 99% Correlation Coefficient = .90

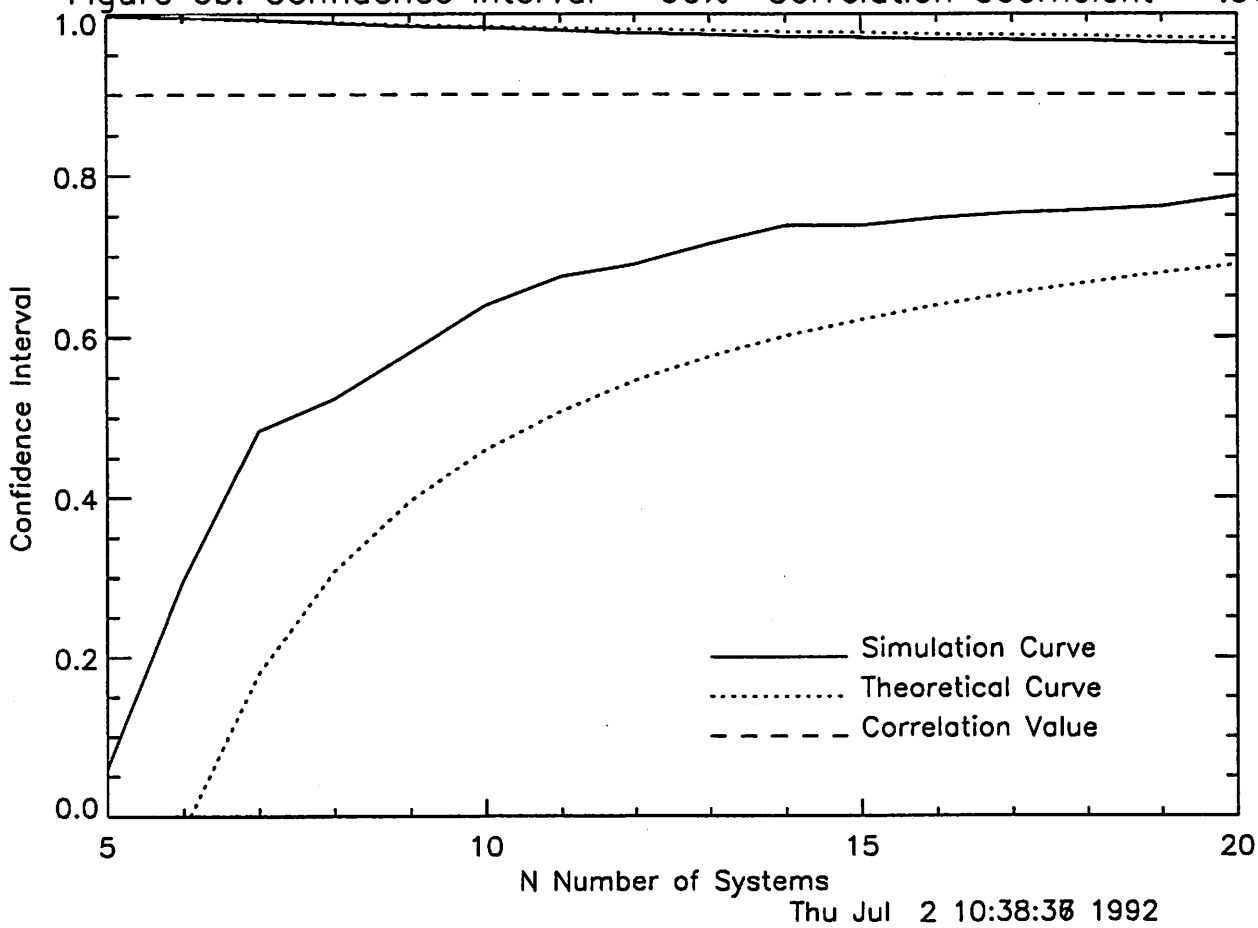

Figure 7a. Confidence Interval = 99% Correlation Coefficient = .95

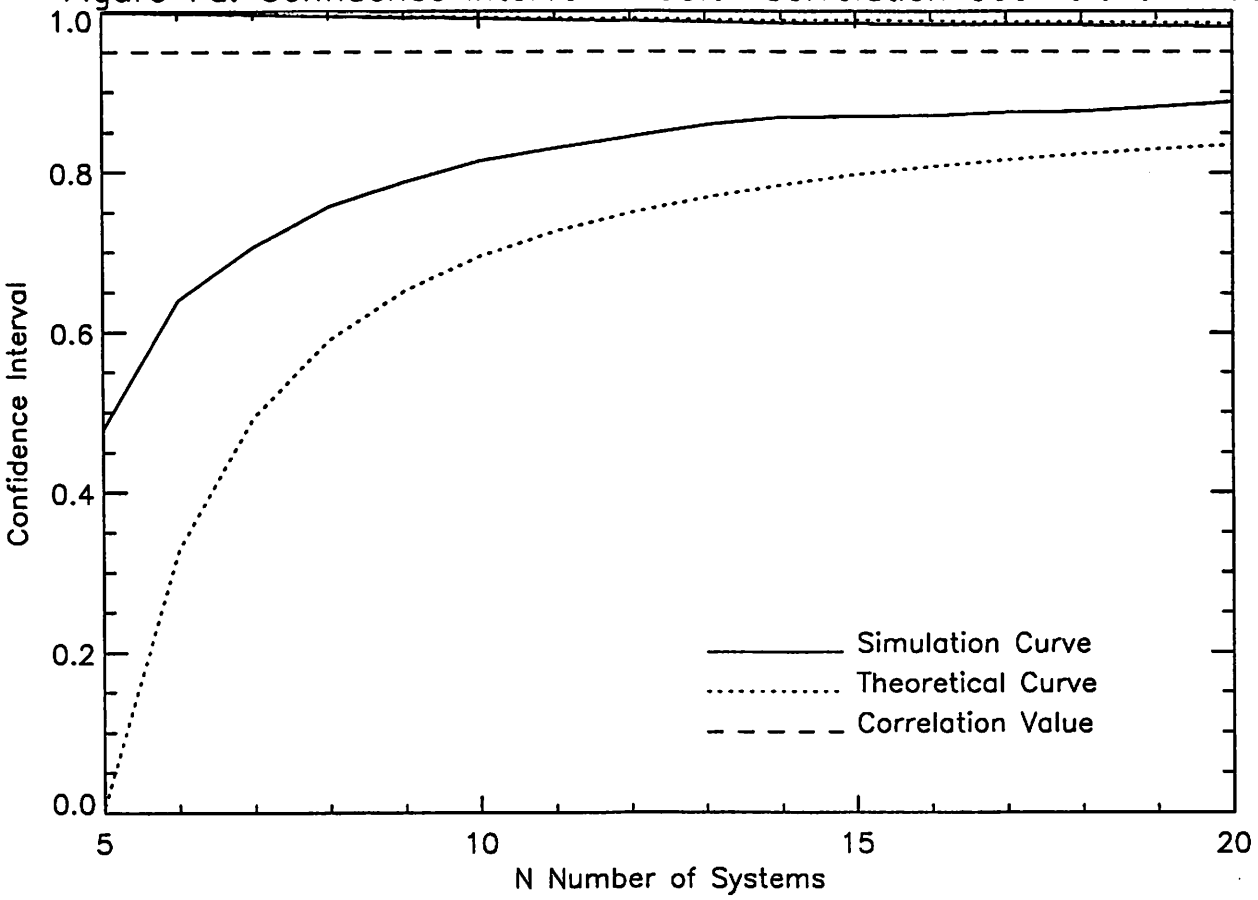

Figure 7b. Confidence Interval = 99% Correlation Coefficient = .99