NOTICE: When government or other drawings, specifications or other data are used for any purpose other than in connection with a definitely related government procurement operation, the U. S. Government thereby incurs no responsibility, nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use or sell any patented invention that may in any way be related thereto.

DEPARTMENT OF THE NAVY

DAVID TAYLOR MODEL BASIN

HYDROMECHANICS

○

AERODYNAMICS

○

STRUCTURAL MECHANICS

○

APPLIED MATHEMATICS

A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM

by

Donn J. Prendergast and Robert E. Dalton

Prepared for the Bureau of Ships.
Distributed only upon their authorization.

21

OPERATIONS RESEARCH DIVISION
APPLIED MATHEMATICS LABORATORY

RESEARCH AND DEVELOPMENT REPORT

January 1962 Report 1557

A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM

by

Donn J. Prendergast and Rcbert E. Dalton

January 1962

Report 1557

TABLE OF CONTENTS

ABSTRACT

The purpose of this report is to permit management to quickly and inexpensively evaluate a real time data processing system and to express a statistical confidence in the validity of their evaluation.

INTRODUCTION

This report results from a conclusion, by the authors, that there presently exist no standards for the manner and procedures by which a real time electronic computer data processing system can be evaluated.

The modern world is in a dynamic era of technological advance. Each program system developed for the processing of data on electronic computers appears to be soon replaced by a larger or "better" system that shows promise of doing the required job faster and better. It is no wonder, then, that an ever increasing burden is placed upon management in the quest to seek the "best" solution to the computer problems that face it. Unfortunately, however, this attempt to select the best solution often involves the choice of computer programs that are far too complex and intricate to be grasped in their entirety. Different language is involved, the sciences are called upon constantly and even the computer people become a "breed unto themselves." More often than not, then, management is called upon to accept as gospel a processing system that they cannot hope to completely understand, much less to evaluate. The added factor that the development of data processing often requires many months of painstaking labor increases management's burden by requiring them to determine the worth of a system long after the original specifications have been written. The net result is the rather ludicrous situation which requires management to approve (or in some cases buy) a system which is too complicated to understand completely without extensive exposure, yet which may contain shortcomings when evaluated in the light of current operational requirements.

Such a situation may or may not be critical, depending on the type of data that is being processed and the requirements of management. Certainly, however, it can be seen that in an area such as the real time processing of strategical and tactical information upon which military commanders will make their decision, an incorrect choice of a system can prove disastrous (or at the very least, embarrassing). It is also evident that regardless of whether the system involves the military or not, the possibility of an incorrect choice is certainly undesirable.

The answer to the problem, as we find it then, is to provide management with the tools by which they can effectively evaluate a proposed (or operating) real time data processing system in a practical manner without the necessity of employing a large group of specialists specifically for this purpose.

Recently a problem of the nature discussed was assigned to the authors; i.e., to evaluate the effectiveness of a large digital computer program (to see if it did in fact do the job which it was supposed to do). It became apparent as the preliminary planning of the evaluation progressed that there was very little written in this area that would provide guidelines by which the evaluation could be accomplished. It is true that much has been written about the theory of evaluation and that similarly many large scale evaluations have been accomplished at great expense and with considerable time involved. While such theory and such evaluations have been of value in the past and will certainly continue to be of value in the future, it would seem that management has a right to expect that they can judge a real time system quickly and easily with the use of relatively inexperienced personnel now that computer processing has become an integral part of their operations. This, then, is the purpose of this report, that is, to permit management to quickly and inexpensively evaluate a real time system and to express a confidence in the validity of their evaluation.

The authors make no claim that the procedures presented herein are necessarily the best nor the only ones available--to the contrary we feel that this document might provide merely a stepping stone upon which further improvements may be made. It is considered, however, that with the factors of cost, time and personnel, the procedures recommended will provide a practical yardstick by which management may make an intelligent decision.

LOGICAL CONSIDERATIONS

The culmination of the development of any large data processing system is the assembly and testing of the prototype system against the specifications for which it is designed. Only when it is assembled and operated as a finished product can the compatibility of the subsystems with one another be assessed realistically. Similarly, it is only when assembled and operated that the compatibility of the program system as a whole with the computer equipment can be assessed realistically. The end result of the evaluation provides the groundwork for the final design and future modification of the system. If the results show that the operational requirements are not fully met, then the requirements must be altered or the system program laid out for redesign to meet the requirements. If the results show that the operational requirements have been met, then the results must be documented to provide a basis for future design change and expansion.

While there exist wide differences in the techniques that may be employed in conducting an evaluation of this type, the final general criteria by which the worth of a data processing system is measured are reasonably consistent. The staff or management personnel who make use of the information produced are interested not so much in the programming steps involved, or the evaluation techniques used, but are more concerned with a determination that the outputs of the system are as follows:

1. Valid: The computations and resultant outputs of the system must provide true and accurate information within given tolerances.

2. Current: The system must update data with sufficient frequency to insure that all computations take into consideration the most recent factors and reflect them in the outputs.

3. Complete: The outputs should provide all the information necessary to permit interpetation and analysis by the user.

4. Accessible: The information desired of the system must be quickly, easily, and directly obtainable to the user without recourse to complicated procedures requiring involved or lengthy training.

5. Readable: All outputs of the system must be presented in a form that is easily understood without intermediate translation or processing.

6. Usable: All output information should be pertinent and necessary to the requirements of the user, eliminating the waste of time and effort in the interpretation of information that is of no interest or value.

The criteria given above are necessarily general, since this procedure is intended to be applicable to most real-time systems. It must be stressed that, in each particular evaluation, it is most important for management to clearly spell out its system output requirements in detail to evaluating personnel. An evaluator should be able to determine whether a selected input has led to erroneous output in all cases on the basis of the criteria for acceptance and rejection which have been furnished him by the management involved; i.e., whether it is "right" or "wrong."

The logical criteria presented here are essentially designed for a final checkout of a system and program using the outputs of the system as a basis for this determination. Therefore it will be assumed that the system machinery has been determined to be reliable by operating personnel and that the operating staff is sufficiently well-versed in their tasks to insure that human error is not significant. Previous experience has indicated that these two assumptions cannot be taken lightly. If the evaluator finds that these sources of system inefficiency can not be ruled out initally, it is suggested that he consult one of the many excellent references which exist, both on machine reliability studies[1] and human work sampling.[2]

The evaluator should insure that detailed logs are kept by the evaluating staff so that, if and when difficulties occur as to the application of the logical criteria to the outputs, they may be readily resolved by consultation with the appropriate level of management. Also, if specifications or interpretations are changed during the course of the evaluation--or even following it--appropriate revisions might possibly be made to the log and figures adjusted. A costly repeat trial run may thereby be avoided.

The evaluator also must know or be given an appropriate time increment on which to base his random observations during the real-time sample. Some estimate as to the frequency of entrance of inputs into the system would be of much help on this question.

Finally, it is suggested that a simulated input representing each type of actual input be entered into the machine prior to actual real-time testing for two reasons:

1. To check that format checks in the program are working properly for all types of input;

[1] References are listed on page 27.

2. To assure that at least one observation is present on each type of input. As will be discussed later, the mean frequency of occurrence of some type (s) of input may be so low as to rule out an occurrence of an input of a certain type during the actual real-time test. In this case, it is advantageous to state that at least one simulated representative input of that type had cleared through the machine.

STATISTICAL CONSIDERATIONS

Though the evaluator is not primarily interested in *measuring* the effect of time upon the processing of inputs, he is most interested in *eliminating* what bias it may contribute to the reliability estimate. Randomization of input selection with respect to time insures that the probability of selection of a particular activity level of the machine and particular input will be the same as the probability of the selection of any other activity level and input. Any systematic effect of time or activity level, if present, is thereby minimized on all types of input processing.

Two separate random samples form the basis for the statistical inferences which the evaluator must make in support of his conclusions. The first sample, or "stage" of the experiment, is made from a large collection of previous input data for which a total analysis would be uneconomical and time-consuming. Since the real time data processing system replaces either a lesser system or hand calculations dealing with similar inputs, a large amount of previous input data can and must be made available to the evaluator for random sampling for representative criteria.

The second, or "actual" sample is made on location during a relatively short continuous real time run, with randomization with respect to real time. For the purpose of establishing validity, it is mandatory to assure that the actual real time sample which one takes for the reliability measurements is representative of the total population of inputs which one might expect to enter the system over an extended period of time. The criteria which are used for this determination of representativeness of the second sample are the frequency estimates of the various types of input over a long, continuous period of operation.

A. FIRST STAGE: ESTIMATION OF FREQUENCY CRITERIA

As stated previously, the first stage of the evaluation establishes, over as long a period as possible, the relative frequency of each input into the system. Considering each input classification separately, each randomly selected input will either belong to this input classification or not. Thus, it is possible to consider the estimates of each input mean frequency as distributed binomially. Average values will be approximately normally distributed with increasing sample size even when the observations are from a skewed distribution of individual items. Thus, we are able to use the normal curve statistics to advantage, regardless of the shape of the distribution from which the samples are drawn. With the establishment of the sample $\bar{x}$ (average value for each input), if we can derive and use some standard deviation measure for a distribution, we can also establish upper and lower confidence limits for this average value. In estimating this average value, it is far better to make a random sample from a long, continuous period of operation than to use all inputs from a block portion of a long period for consideration, since biases or peaks toward a predominence of one or more inputs are more likely to remain undetected during shorter period of operation. If cost and time allow, a complete--rather than random--selection of inputs over an extended period of time should be even more satisfactory for estimation of the frequencies and associated confidence limits at this stage, if the evaluator is quite sure that the observations are independent.

For clarity, the following hypothetical example is presented: analysis of a real-time data processing system discloses five distinct types of input, A, B, C, D, and E. One thousand inputs of these types are selected randomly from a 5-week period. The results of this sampling are shown in the following table:

Input	Week					Totals
	1	2	3	4	5	
A	32	44	29	40	33	178
B	98	85	100	83	99	465
C	46	38	46	60	33	223
D	25	27	23	15	33	123
E	0	2	5	1	3	11
Total	201	196	203	199	201	1000

The sample relative frequencies are:

$$\bar{a} = 0.178$$

$$\bar{b} = 0.465$$

$$\bar{c} = 0.223$$

$$\bar{d} = 0.123$$

$$\bar{e} = 0.011$$

Though these are the best estimates available, they are at best sample values and it is wise to gain some idea of how accurate the estimation actually is. Either exact v^2 confidence limits[*] or the approximation provided by Student's t-distribution[**] could be used for this determination. For samples of this size and larger, the approximation based on the t-distribution appears quite satisfactory and is far easier to compute. These limits, which can be found in most work sampling and statistical texts, are

[*] See Reference 3, page 698.

[**] See Reference 4, page 77.

$$\bar{\theta} \quad = \quad \bar{x} \quad + \quad t_{\alpha}\sqrt{\frac{\bar{x}\ (1 - \bar{x})}{n - 1}}$$

$$\underline{\theta} \quad = \quad \bar{x} \quad - \quad t_{\alpha}\sqrt{\frac{\bar{x}\ (1 - \bar{x})}{n - 1}}$$

where

n = total number of observations in the sample

α = desired critical region for 100 $(1 - \alpha)$ % confidence limits

$\bar{x}$ = sample mean

t = Student's t with $(n - 1)$ degrees of freedom

The particular confidence level (defined by α) which is desired will probably differ with each particular experiment. It is recommended that an α less than or equal to 0.05 be used.

For 95 percent confidence limits on the true mean frequency θ_A of input A we have ($\alpha = 0.05$)

$$\bar{\theta}_A \quad = \quad \bar{a} \quad + \quad t_{0.05}\sqrt{\frac{\bar{a}\ (1 - \bar{a})}{999}}$$

$$\bar{\theta}_A \quad = \quad 0.178 \quad + \quad (1.96) \quad \sqrt{\frac{(0.178)\ (0.822)}{999}}$$

$$\bar{\theta}_A \quad = \quad 0.178 \quad + \quad (1.96) \quad \sqrt{0.00014646}$$

$$\bar{\theta}_A \quad = \quad 0.178 \quad + \quad (1.96) \quad (0.013002) \quad = \quad 0.203$$

Similarly,

$$\underline{\theta}_A = 0.178 - (1.96)\ (0.013002) = 0.154$$

So that, in repeated sampling,

$$P\ (\underline{\theta}_A < \theta_A < \bar{\theta}_A) = P(0.154 < \theta_A < 0.203) = 0.95$$

Similar calculations for the remaining input types yield

$$\begin{array}{ll}
\bar{\theta}_B = 0.496 & \underline{\theta}_B = 0.434 \\
\bar{\theta}_C = 0.249 & \underline{\theta}_C = 0.197 \\
\bar{\theta}_D = 0.143 & \underline{\theta}_D = 0.103 \\
\bar{\theta}_E = 0.017 & \underline{\theta}_E = 0.005
\end{array}$$

Since the maximum value for $\bar{x}\ (1 - \bar{x})$ is 1/4 (when $\bar{x} = 1/2$), we see that the larger deviations about a sample frequency will occur when the frequency approaches 1/2 and that smaller deviations will occur when the frequency is near 0 or 1. Also, since the limits are inversely proportional to $\sqrt{n}$, they will be closer together (or the estimates more accurate) as the sample size n is increased. For this reason, the largest possible sample must be taken in the first stage.

B. SECOND STAGE: REPRESENTATION AND RELIABILITY

With estimates of frequencies and associated confidence limits determined, the actual real time sample must be devised for the determination of the reliability of the system. The organization of this second stage of sampling consists of the following steps:

1. A determination of the sample size necessary.

2. A randomization of input selection with respect to real time (or activity level of the system).

3. When the sample is accepted as representative*, a determination of the reliability of the total system and--if possible--each type of input, with appropriate confidence limits.

Considered together, the frequencies from sample 1 are estimates of the actual mean frequencies θ_i of the multinomial distribution with frequency function

$$p(x_A, x_B, \ldots, x_K) = \frac{n!}{x_A!\, x_B! \ldots x_K!} \theta_A^{x_A} \theta_B^{x_B} \ldots \theta_K^{x_K}$$

This formula denotes the probability that input A will occur x_A times, input B will occur x_B times, ... , and input K will occur x_K times in n observations of the second sample, i. e.,

$$x_A + x_B + \ldots + x_K = n$$

Further, it is assumed that

1). the theoretical mean frequencies θ_i sum to one;

2). the input types are mutually exclusive; and

3). the second sample observations are stochastically independent.

This latter assumption usually necessitates a random selection of the inputs to assure that each single input has an equal change of selection for the second sample estimates of input type frequency.

Our criterion for representativeness is the Chi-square "Goodness of Fit" test. It is well-suited for the multinomial distribution, as it affords a single test for the representativeness of all frequency estimates. The formula is

*When the second stage sample is rejected as unrepresentative, the experiment reverts to the first stage. A more recent first sample should be taken, of a larger size if possible. Frequency estimates and confidence limits must be reformulated for the input types, since the rejected sample furnishes us with significant evidence that the relative frequencies θ_i are changing. A close comparison of the frequencies obtained in sample 2 and the confidence limits established in sample 1 should give a good indication of <u>which</u> input frequencies are changing.

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n\theta_i)^2}{n\theta_i} \quad ; \quad f = k - 1$$

where

k = number of input typ.s

θ_i = theoretical frequency of occurrence of input i as estimated in sample 1

x_i = number of occurrences of input i in sample 2

f = degrees of freedom

n = number of observations in sample 2

Examination of the derivation of this formula* reveals that the sampling distribution of χ^2 tends to a limiting distribution independent of the probability function (in this case the multinomial), depending merely on the parameters (in this case the θ_i) which are to be estimated from the sample.

Since this test employs the normal approximation, it is necessary to have a second sample size n large enough for such an approximation to be valid. Fisher recommends that n be large enough so that each $n\theta_i > 5$ ** Therefore, the minimum sample size is obtained by taking the smallest frequency from sample 1, θ_s, and letting

$$n = 5 / \theta_s$$

Noting that n must be an integer, round off to the next highest integral value.

Referring to the hypothetical problem again, determine the random sample size necessary to validate the normal approximation for sample 2. The smallest input frequency is

$$\theta_s = \theta_E = 0.011$$

Solving for n,

$$n = 5 / 0.011 = 454.545$$

*See Reference 5, page 417.

**See Reference 4, page 135.

Rounding off to the next higher integer value, n ≅ 455. Thus, 455 random observations are necessary in sample 2 to assure the validity of the Chi-squared "Goodness of Fit" test for representativeness. On the basis of the number of observations necessary for the sample, the randomization on time increments can be derived as shown in Section IV on evaluation procedure.

We illustrate the use of the Chi-squared test by the following example: Suppose 455 inputs are randomly selected in time, yielding the results illustrated in the following table.

Input	Number (x_i)	Frequency (sample 2)
A	87	0.191
B	220	0.484
C	92	0.202
D	53	0.116
E	3.	0.007

Since there are five input types, we compute a Chi-square with four degrees of freedom. In this example

$$\chi^2 = \frac{(x_A - (455)\,\theta_A)^2}{(455)\,\theta_A} + \frac{(x_B - (455)\,\theta_B)^2}{(455)\,\theta_B} + \frac{(x_C - (455)\,\theta_C)^2}{(455)\,\theta_C}$$

$$+ \frac{(x_D - (455)\,\theta_D)^2}{(455)\,\theta_D} + \frac{(x_E - (455)\,\theta_E)^2}{(455)\,\theta_E}$$

$$= \frac{(87 - (455)\ (0.178)\ )^2}{(455)\ (0.178)} + \frac{(220 - (455)\ (0.465)\ )^2}{(455)\ (0.465)}$$

$$+ \frac{(92 - (455)\ (0.223)\ )^2}{(455)\ (0.223)} + \frac{(53 - (455)\ (0.123)\ )^2}{(455)\ (0.123)} + \frac{(3 - (455)\ (0.11)\ )^2}{(455)\ (0.011)}$$

$$= 2.6247$$

Since $\chi^2_{0.40} = 2.75$ ($P\{\chi^2 < 2.75\} = 0.40$) with four degrees of freedom, the value obtained from the sample is in the lower 40% of the Chi-squared distribution for four degrees of freedom. Since we may expect values greater than this more than 60% of the time, there are no grounds present for rejecting the hypothesis that sample two is from the same population estimated by sample one. If the sample Chi-squared value were greater than $\chi^2_{0.95} = 9.49$ for four degrees of freedom, we would have significant evidence for rejecting the sample as unrepresentative.

Having established the representativeness of the sample, the final step of the evaluation is to obtain a reliability estimate. The reliability measurements are made by following the selected inputs through the machine processing and by analyzing the outputs resulting from the selected random inputs on the basis of the logical criteria. Since the outputs resulting from a given input are either correct or in error, we have a binomial distribution. The probability that there are not more than x erroneous outputs resulting from a total of n inputs is therefore given by the cumulative binomial distribution. This is known to be equivalent to an incomplete beta integral, as is the v^2 or Fisher's F distribution[*], which is widely tabulated. Since our second sample will necessarily be smaller than the first, exact confidence limits are used rather than those afforded by the approximation with the t distribution. The limits on the actual frequency of error of the population of inputs from which we sample are based on the v^2 or Fisher's F distribution and are[**]

$$\bar{E} = \frac{(x_0 + 1)\, v^2{}_{1-P_1}(f_1, f_2)}{n - x_0 + (x_0 + 1)\, v^2{}_{1-P_1}(f_1, f_2)}$$

$$f_1 = 2(x_0 + 1)$$

$$f_2 = 2(n - x_0)$$

[*] See Reference 3, pages 672-675.

[**] See Reference 3, page 698.

and

$$\underline{E} = \frac{x_0}{x_0 + (n - x_0 + 1)\, v^2_{P_2}\ (f_1^{*},\ f_2^{*})}$$

$$f_1^{*} = 2\ (n - x_0 + 1)$$

$$f_2^{*} = 2x_0$$

where

n = total observations in sample 2

x_0= total number of inputs resulting error in sample 2

$P_2 - P_1 = 1 - \alpha$, which defines $100(1 - \alpha)\%$ confidence limits

$P_2 = P\ \{ x < v^2_{P_2}\ (f_1^{*},\ f_2^{*})\}$

$P_1 = P\ \{ x < v^2_{P_1}\ (f_1,\ f_2)\}$

where x is a random variable from 0 to ∞

Assume now that the 455 inputs selected in the example have been analyzed with respect to their effect on the system as pictured by the outputs which are influenced by them. The results may be summarized in the following table:

Input	Number (x_i)	Number Correct	Number in Error (x_0)	E*	R**
A	87	85	2	0.0230	0.9770
B	220	220	0	0.0000	1.0000
C	92	91	1	0.0109	0.9891
D	53	49	4	0.0755	0.9245
E	3	3	0	0.0000	1.0000
Overall	455	448	7	0.0154	0.9846

*E = frequency of error (as estimated by sample 2)

**R = (1 - E) reliability (as estimated by sample 2)

Ninety-five percent confidence limits on the overall error frequency are as follows: $(P_2 = 0.975,\ P_1 = 0.025)$

$$\bar{E} = \frac{(7+1)\, v^2_{0.975}\,(16{,}896)}{448 + (7+1)\, v^2_{0.975}\,(16{,}896)}$$

$$= \frac{8(1.82)}{448 + 8(1.82)} = 0.0315$$

$$\underline{E} = \frac{7}{7 + (449)\, v^2_{\ 0.975}\,(898{,}14)} = \frac{7}{7 + (449)\,(2.49)} = 0.0062$$

Thus, the true overall reliability is between

$$\bar{R} = 1 - \underline{E} = 0.9938$$

and $\underline{R} = 1 - \bar{E} = 0.9685$

in 95% of all samples of this size. It is well in general to obtain a high percentage of confidence, but the reliability confidence limits $\bar{R}$ and $\underline{R}$ will be closer together at the lower confidence percentages. This should be remembered if the confidence interval for the evaluation is required to be within a certain numerical tolerance, regardless of the confidence percentage.

Limits may be computed for each input type if desired. For this example, one has the various frequencies of error with 95% confidence limits as follows:

$$\begin{array}{llll}
\bar{E}_A = 0.0852 & \qquad & \underline{E}_A = 0.0028 \\
\bar{E}_B = 0.0166 & \qquad & \underline{E}_B = 0.0000 \\
\bar{E}_C = 0.0591 & \qquad & \underline{E}_C = 0.0003 \\
\bar{E}_D = 0.1820 & \qquad & \underline{E}_D = 0.0209 \\
\bar{E}_E = 0.7076 & \qquad & \underline{E}_E = 0.0000
\end{array}$$

Several items may be noted concerning the preceding limits. Note that, besides offering a more exact derivation of confidence limits for the binomial distribution, these limits also offer a convenient expression for confidence limits on frequency of error E *even when no errors actually occur* during the actual evaluation. $\underline{E}$ is immediately seen to equal zero when $x_o = 0$, but examination discloses that $\bar{E}$ is never zero. If the normal approximation or t-test approximation were used in this case no limits could be drawn at all.

Also, the role of the number of observations of each input is clearly seen in these limits, as only the limits for inputs A-C appear to be of any use. Inputs D and E have so few occurrences as to make the limits too far apart to be useful at the 95% level. What is important in the foregoing calculations is that the *overall* reliability may be estimated within fairly close bounds even if some of the types of inputs appear quite infrequently during the second sample. Although it is advisable to gain an actual case of the occurrence of each type input during the actual real time sampling, the absence of one or more types does not at all negate the drawing of confidence limits for the reliability estimate.

It is conceivable that in many instances no errors at all will occur during the real time evaluation (sample two). It is well to note that confidence limits can be drawn for this very desirable phenomenon also. In our example, with no errors in 455 inputs, the upper 95% confidence limit on the frequency of error is

$$\bar{E} = \frac{v^2_{0.975}\,(2,910)}{455 + v^2_{0.975}\,(2,910)}$$

$$= \frac{3.70}{458.70} = 0.0081$$

The lower limit $\underline{E}$ is, of course, zero. This means that the evaluator could state with a 95% confidence that the overall reliability of the system would be between

$$\bar{R} = 1 - \underline{E} = 1.0000 \qquad \text{and} \qquad \underline{R} = 1 - \bar{E} = 0.9919$$

in repeated sampling on the basis of this sample.

THE CONDUCT OF THE EVALUATION

The following procedural steps present the chronological order by which the evaluation should take place and will serve also as a suggested outline for the final report. In summary, both the evaluation and the final report of that evaluation consist of the following basic steps:

A. Preliminary orientation with the system to be evaluated.

STAGE ONE

B. Assembly of a large continuous segment of previous inputs.

C. Estimation of input frequencies with associated confidence limits.

STAGE TWO

D. Random sampling of actual inputs during actual operation on the basis of real time.

E. Establishing the representativeness of the real time random sample.

F. Estimation of system reliability with associated confidence limits.

The purpose of this section is to permit the conduct of the evaluation with relatively inexperienced personnel; accordingly, an attempt has been made to reduce the material contained herein to its simplest state. Unfortunately, however, the results of the evaluation are based on certain statistical theory, and it is stressed that each and every step that is indicated be strictly adhered to. The formulas that have been used should require merely substitution therein and do not necessarily presuppose extensive mathematical knowledge. If the formulas are not clear, however, they should be clarified.

A breakdown of the foregoing basic steps results in the following evaluation procedure:

A. PRELIMINARIES

Step 1. Broad Statement of System:

Reduce to written form a broad summary of the system to be evaluated. This may be done by a condensation of the specifications combined with a brief description of how the system is designed to meet these specifications.

Step 2. Detailed Statement of System with Regard to Types of Inputs and Outputs:

Though the evaluator should have a general picture of the flow of information through the system, the statistical nature of the evaluation makes it imperative that he be especially familiar with the types, formats, and tolerances involved in the inputs and outputs; and that he demonstrate this familiarity in the final report.

Step 3. A Meeting with Operating Personnel to Determine in Detail what Satisfies Logical Criteria for Output:

Here the general criteria presented in the logical section should become concrete to the evaluator. He should be able to tell whether or not any selected output satisfies the six general logical categories: Validity, Currentness, Completeness, Accessibility, Readability, and Usability. From discussions with operating personnel, he should be able to compare original specifications against current operational requirements and discuss any discrepancies in the final report.

Step 4. Establishment of Input Categories:

All of the inputs to the system must be firmly typed and categorized. In most cases this will have been accomplished by the programming group in their assignment of an individual identification code to each input. If this has not been done, then the evaluator should assign his own identification to remove the possibility of ambiguity.

STAGE ONE

B. ASSEMBLY OF A LARGE CONTINUOUS SEGMENT OF PREVIOUS INPUTS

Step 5. Assemble Largest Continuous Segment of Previous Inputs Possible:

The main stress here is on the word "continuous." In the first stage sample, time is only important insofar as there are no breaks in time present during the period when the inputs are compiled. The total inputs may be arranged in any arbitrary order for random sampling, keeping in mind that the greater the number of observations, the more accurate the frequency estimates for the various types of input.

Step 6. Random Sampling - Stage One:

The use of the random sampling numbers* in this sample is outlined as follows:

a. Assemble inputs in any order.

b. Beginning at the upper left-hand corner of a two-figure random number table, count this many inputs and select the last one counted. Remove this input and note its type.

c. Select the next number below and proceed as above, column by column.

d. When the end of the inputs is reached, begin at start again; continue as long as possible.

e. Total tallies for each input type and for entire sample.

C. ESTIMATION OF INPUT FREQUENCIES WITH ASSOCIATED CONFIDENCE LIMITS

Step 7. Use of Formulas for Stage One:

a. The sample relative frequency for each type input is obtained by taking the ratio of the total number of inputs of a certain type to the total number of all inputs in the random sample. For example, in the hypothetical problem presented in the statistical section of this report, input A occurred 178 times in 1000 inputs. Thus, the sample relative frequency for input A is

$$\bar{a} = \frac{178}{1000} = 0.178$$

b. Confidence limits

$$\bar{\theta} = \bar{x} + t_{\alpha} \sqrt{\frac{\bar{x}\,(1 - \bar{x})}{n - 1}}$$

$$\underline{\theta} = \bar{x} - t_{\alpha} \sqrt{\frac{\bar{x}\,(1 - \bar{x})}{n - 1}}$$

*See Reference 6, pages 92-97.

In the formulas for $\bar{\theta}$, the upper confidence limit, and $\underline{\theta}$, the lower confidence limit, each input must be considered separately and $\bar{x}$ represents whatever particular sample relative frequency with which we are concerned; e.g., $\bar{x}$ is $\bar{a}$ when we are discussing input A. The t_α value is obtained from Hald's Table IV[*] by reading (n - 1), one less than the total number of inputs in the sample, on the vertical scale; and $\alpha = 2\ (1 - P)$, where P is the desired percentage of probability on the bottom scale.

For 90% confidence limits, use $\alpha = 10$ (%) (1.645)
95% confidence limits, use $\alpha = 5$ (%) (1.960)
98% confidence limits, use $\alpha = 2$ (%) (2.326)
99% confidence limits, use $\alpha = 1$ (%) (2.576)
99.8% confidence limits, use $\alpha = 0.2$(%) (3.090)
99.9% confidence limits, use $\alpha = 0.1$(%) (3.291)

Values given in parentheses may be used accurately for t-values for samples of 1000 or more. Linear interpolation will suffice for intermediate values on the vertical scale.

c. It is statistically important for the desired percentage of confidence to be selected _prior_ to the determination of the sample relative frequencies.

d. Convenient tables[**] are available for the determination of the square root term in the formulas for $\bar{\theta}$ and $\underline{\theta}$.

With the input frequencies and associated confidence limits obtained, the first stage sampling comes to a close. It now becomes the responsibility of the evaluator to rule out certain causes of error in the actual machine run so that the Stage Two sample will properly reflect the reliability of the program versus specifications and/or operational requirements on the basis of the logical criteria. This elimination of undesired sources of error consists of three steps.

[*] See Reference 6, page 39.

[**] See Reference 6, pages 84-87.

Step 8. Determination of Reliability of Machinery and Operation:

The machinery must exhibit a previous percentage of downtime small enough for the evaluator to be reasonably sure that a breakdown will not occur during the period required for the actual real time sample. Operating personnel should be so well-versed in their duties that no appreciable errors in input preparation or output representation will occur during the Second Stage sampling. If it appears that these factors must be tested, procedures for such evaluation exist[1,2] and should be consulted.

Step 9. Test Each Type Input, Format Checks, and Tolerances:

a. Run one example of each type input through the System to determine that at least one example of each type input is acceptable to the system and results in a correct output.

b. All checks for proper format and tolerances must be checked for possible programming errors before beginning the actual evaluation run.

Step 10. Time Allowance for Getting System Underway:

Following machinery downtime or operational checkout, real time systems often require a period of time to reach normal capacity and operation. Since the distribution of input data during this period may be highly unrepresentative of normal operating conditions, it is well to postpone the initiation of random input selection until this period has been passed sucessfully and the system is operating "normally". The evaluator should determine if such a condition exists for the particular system with which he is involved by personal familiarity or consultation with operating personnel. Since such a condition will result whenever a machinery shutdown or major operational error occurs, the actual real time sampling should be started over again after such an occurrence.

STAGE TWO

D. RANDOM SAMPLING OF ACTUAL INPUTS DURING ACTUAL OPERATION ON THE BASIS OF REAL TIME

Step 11. Determine the Minimum Necessary Size for the Second Sample:

a. Select the input with the smallest frequency as determined in Step 7a.

b. Divide this frequency into 5.

c. Round the resultant quotient off to the next highest integer; this integer is n, the minimum sample size for sample two.* (If $n < 100$, set $n = 100$.)

With the minimum sample size thus determined, we can now proceed to distribute these sample observations over a period of actual running time in a random fashion.

Step 12. Random Sampling on the Basis of Time - Stage Two:

a. Deciding on a Time Increment. For this procedure, it is most desirable to have some estimate as to how often inputs will enter the system, so that randomization can be based on the average length of system "waiting" time between inputs. This estimate may be obtained from the times of previous inputs used in Stage one sampling or may be proposed by operating personnel. By the use of the random number table, we are assured that the observations used in this second sample will be random; picking a realistic increment aids in attaining the shortest possible sampling period necessary to obtain the number of actual observations required. Periodic display or other output characteristics may necessitate the selection of a more inefficient increment on which to base the random observations. The important points here are that an increment be selected and that it be as close as possible to the average length of system waiting time.

b. Randomization. Suppose the increment decided on was 6 minutes. By placing a decimal point before each number of the random number table, we have immediately 7500 random fractions between 0 and 1, with mean approximately 0.50. Multiplying each of these numbers by 2 times 6, or 12, as they

*Of course it is possible here to determine a minimum sample size that is actually larger than the original extended sample. A very small input frequency for a certain type will do this. In this case two alternatives are open to the evaluator:

1. He may choose to drop the very infrequent input from consideration of the minimum second sample size. In this case, he must also drop it from use in the Chi-square "Goodness of Fit" test also. This alternative should be used when the next most infrequent input leads to a considerably more economical sample size.

2. He may choose to pool several very small frequencies into one input class and use the total group frequency (if it is still the smallest) for the determination of n. In this case, it is most important that this group identification be maintained during the subsequent "Goodness of Fit" test.

are selected will merely change the table to 7500 random numbers between 0 and 12, and setting the mean approximately at 6, which was the predetermined waiting time. We follow the steps presented in Stage One for selection of the random numbers and multiply each number by twice the average waiting time to determine the times at which observations will be made. Using the same table as in Stage One, we assume the actual sample begins at 1200. We round off to the ne^ ·est minute.

First number selected: 15
12 x 0.15 = 1.80. Round off to 2
Select first input entered into system after 1202.

Second number selected: 85
12 x 0.85 = 10.20. Round off to 10
Select next input entered into system 10 minutes after first.

Third number selected: 47
12 x 0.47 = 5.64. Round off to 6
Select next input entered into system 6 minutes after second.

Etc.

Proceed from latest time calculated. Continue until minimum number of observations is compiled. This number of observations is the "n" obtained in Step 11.

E. ESTABLISHING THE REPRESENTATIVENESS OF THE REAL TIME RANDOM SAMPLE

Step 13. <u>Determination of Representativeness:</u>

The following procedure concerns the use of the "Goodness of Fit" formula*

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - n\theta_i)^2}{n\theta_i} \qquad f = k - 1$$

*See Reference 6, pages 40-43.

a. Use the value obtained in Step 11 for n.

b. Use estimates obtained in Step 7a for θ_1 to θ_k.

c. Count numbers of inputs of each type in sample two. These numbers are x_1 to x_k.

d. k is the number of types of input. (If the modifications mentioned in footnote of Step 11, Stage Two, are used, k will be reduced accordingly.) Sum as indicated.

e. Select a confidence interval. Usual percentages are greater than 0.95.

f. Find f on vertical scale.

g. Check Chi-square value for desired percentage and given f. Tf it is larger than the number computed by the formula, accept the sample as representative; if it is smaller, reject. If sample two is rejected as unrepresentative, the evaluation reverts to Step 5 as a new large sample must be taken preferably with more recent data.

h. Prior to reverting to Step 5, it is recommended the $\frac{x_i}{n}$ values be compared with their respective $\bar{\theta}$ and $\underline{\theta}$ values determined in the first stage, for by so doing it might be possible to determine which input (s) caused the discrepancy which in turn might indicate the existence of a system change, etc.

F. ESTIMATION OF SYSTEM RELIABILITY WITH ASSOCIATED CONFIDENCE LIMITS

Step 14. <u>Use of Reliability Formulas:</u>

The following steps concern the use of the formulas

$$\bar{E} = \frac{(x_0 + 1) \ v^2{}_{1-P_1} \ (f_1, f_2)}{n - x_0 + (x_0 + 1) \ v^2{}_{1-P_1} (f_1, f_2)}$$

$$f_1 = 2 \ (x_0 + 1)$$

$$f_2 = 2 \ (n - x_0)$$

and

$$\underline{\underline{E}} = \frac{x_0}{x_0 + (n - x_0 + 1)\ v^2_{P_2}\ (f_1^*, f_2^*)}$$

$$f_1^* = 2\ (n - x_0 + 1)$$

$$f_2^* = 2x_0$$

a. Use the number n computed in Step 11 for overall reliability determination.

b. Count total number of inputs resulting in error. (This is not necessarily the same as the total number of <u>outputs</u> in error, since one input may lead to several erroneous outputs. Nor is it the number of inputs which affect a given erroneous output, since only one or a portion of the inputs leading to that output may have led to the error). This is x_0 for the overall reliability estimate.

c. Count the number of inputs <u>of each type</u>. These are the n's for reliability estimate for <u>each type</u> input.

d. Count number of inputs <u>of each type</u> resulting in erroneous output. These are the x_0's for reliability estimates <u>for each type</u> input.

e. Usual values for P_1 and P_2 are as follows:

90% confidence:	$P_1 = 0.05$	$P_2 = 0.95$
95% confidence:	$P_1 = 0.025$	$P_2 = 0.975$
98% confidence:	$P_1 = 0.01$	$P_2 = 0.99$
99% confidence:	$P_1 = 0.005$	$P_2 = 0.995$

f. Values of v^2 are obtained from Hald's Tables.* The values of $(1 - P_1)$ or P_2 respectively will determine which v^2 table to use; f_1 is entered horizontally and f_2, vertically. Be sure that f_1 and f_2 are not interchanged since $v^2\ (f_1, f_2)$ does not usually equal $v^2\ (f_2, f_1)$.

*See Reference 6, pages 47-59.

g. $\overline{E}$ and $\underline{E}$ are confidence limits on the frequency of error on the basis of the second sample. To obtain the confidence limits for reliability,

$\overline{R} = 1 - \underline{E}$ = upper confidence limit

$\underline{R} = 1 - \overline{E}$ = lower confidence limit

h. As shown in the hypothetical example in the mathematics section, reliability estimates for a particular type input are usually not very accurate unless over 75 inputs of a certain type have been selected in the random sample.

REFERENCES

1. Chorafas, D.N. "Statistical Processes and Reliability Engineering," D. Van Nostrand, Princeton, New Jersey (1960).

2. Heiland, R.E. and Richardson, W.J. "Work Sampling," McGraw-Hill Book Co., New York (1957).

3. Hald, A. "Statistical Theory with Engineering Applications," John Wiley & Sons, New York (1952).

4. Anderson, R.L. and Bancroft, T.A. "Statistical Theory in Research," McGraw-Hill Book Co., New York (1952).

5. Cramer, H. "Mathematical Methods of Statistics, "Princeton University Press (1946).

6. Hald, A. "Statistical Tables and Formulas," John Wiley & Sons, New York (1952).

BIBLIOGRAPHY

1. Cochran, W.G. and Cox, G.M. "Experimental Designs," Second Edition, John Wiley & Sons, New York (1957).

2. Feller, W. "An Introduction to Probability Theory and Its Applications," Second Edition, John Wiley & Sons, New York (1957).

3. Mood, A.M. "Introduction to the Theory of Statistics," McGraw-Hill Book Co., New York (1950).

INITIAL DISTRIBUTION

Copies	
5	CHBUSHIPS
	1 Tech Info Br (Code 335)
	1 Computer Sys & Appl (Code 732)
	1 OP Con Cntr Proj Off (Code 671B)
	1 Spcl Faclts & Sys Eng Br (Code 680)
	1 Comm & Computer Sys Br (Code 686)
4	CHONR
	1 Res Dir (Code 402)
	1 Dir, Naval Analysis (Code 405)
	1 Dir Math Sci Div (Code 430)
	1 Dir, Sys Analysis (Code 492)
3	CHONO
	1 Comm Sys Program Div (Op-74)
	1 Asst for Naval Comm Sys (Op-94E)
	1 Oper Analysis & Data Proc Br (Op-335)
1	BUSANDA
	1 Res & Dev (Code OW)
10	ASTIA
2	CINCLANTFLT
2	Fleet Oper Control Cntr, LANTFLT
3	CINCPACFLT
2	Fleet Oper Control Cntr, PAC
5	CINCPAC (Code J02C)
1	JCS
	I Cmmd & Control Dev Group
1	CINCPACAF
1	CINCUSARPAC
1	COMASDEFORPAC
1	COMHAWSEAFRON
2	COMWESTSEAFRON
1	COMBARLANT
1	CO & DIR NEL, San Diego, Calif.
1	Dr. J. W. Gebhard c/o Bureau of Naval Weapons Rep 8621 Georgia Avenue, Silver Spring, Md.

David Taylor Model Basin. Report 1557.
A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM, by Donn J. Prendergast and Robert E. Dalton. Jan 1962. ii, 29p. illus., tables, refs. (Prepared for the Bureau of Ships. Distributed only upon their authorization)
UNCLASSIFIED

The purpose of this report is to permit management to quickly and inexpensively evaluate a real time data processing system and to express a statistical confidence in the validity of their evaluation.

1. Data processing systems-Effectiveness
I. Prendergast, Donn J.
II. Dalton, Robert E.

David Taylor Model Basin. Report 1557.
A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM, by Donn J. Prendergast and Robert E. Dalton. Jan 1962. ii, 29p. illus., tables, refs. (Prepared for the Bureau of Ships. Distributed only upon their authorization)
UNCLASSIFIED

The purpose of this report is to permit management to quickly and inexpensively evaluate a real time data processing system and to express a statistical confidence in the validity of their evaluation.

1. Data processing systems--Effectiveness
I. Prendergast, Donn J.
II. Dalton, Robert E.

David Taylor Model Basin. Report 1557.
A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM, by Donn J. Prendergast and Robert E. Dalton. Jan 1962. ii, 29p. illus., tables, refs. (Prepared for the Bureau of Ships. Distributed only upon their authorization)
UNCLASSIFIED

The purpose of this report is to permit management to quickly and inexpensively evaluate a real time data processing system and to express a statistical confidence in the validity of their evaluation.

1. Data processing systems-Effectiveness
I. Prendergast, Donn J.
II Dalton, Robert E.

David Taylor Model Basin. Report 1557.
A SCIENTIFIC EVALUATION OF A REAL TIME DATA PROCESSING SYSTEM, by Donn J. Prendergast and Robert E. Dalton. Jan 1962. ii, 29p. illus., tables, refs. (Prepared for the Bureau of Ships. Distributed only upon their authorization)
UNCLASSIFIED

The purpose of this report is to permit management to quickly and inexpensively evaluate a real time data processing system and to express a statistical confidence in the validity of their evaluation.

1. Data processing systems-Effectiveness
I. Prendergast, Donn J.
II. Dalton, Robert E.