FOREWORD

This is the final report on research on a systems reliability model. It discusses the role of the human-initiated malfunction in weapon systems research and development testing. The research was conducted under Contract No. AF 33(616)-7968, Project No. 7184, "Human Performance in Advanced Systems," Task No. 718404, "Advanced Systems Human Engineering Design Criteria," between March 1961 and August 1962. The project is administered for the 6570th Aerospace Medical Research Laboratories by Charles Bates, Jr. Mr. Walter J. Huebner, Performance Requirements Section, Human Engineering Branch, Behavioral Sciences Laboratory, served as contract monitor. At Battelle Memorial Institute, the research is performed under the administrative direction of John K. Wetherbee, Chief, Systems Engineering Division.

ABSTRACT

This research project was concerned with determining whether a system's performance conforms to its design objectives from consideration of failure data, including human-initiated malfunctions collected during development testing. In addition, the usefulness of this data for predicting the system's reliability was examined. The report discusses the interaction of human factors activities with other developmental testing activities. The relationship of the human-initiated malfunction to the reliability growth in development is examined. A method for monitoring performance and reliability during development is presented as an aid to determining whether or not a system conforms to its design objectives.

PUBLICATION REVIEW

This technical documentary report has been reviewed and is approved.

Walter F. Grether

WALTER F. GRETHER
Technical Director
Behavioral Sciences Laboratory

TABLE OF CONTENTS

Page

CHAPTER 1

CHAPTER 2

CHAPTER 3

TABLE OF CONTENTS
(Continued)

LIST OF FIGURES

LIST OF TABLES

HUMAN-INITIATED MALFUNCTIONS AND SYSTEM PERFORMANCE EVALUATION

by

Roger C. Van Buskirk and Walter J. Huebner

CHAPTER 1

INTRODUCTION

Purpose of the Research

The purpose of the research program reported herein is to develop a model which can be used first, as an aid for determining whether a system* conforms to design objectives, and second for predicting the system's reliability. This model is concerned primarily with the research and development testing phases of the system's development and not the system's production phase nor its operational status. The primary data source for the model would be malfunction and failure data obtained during the testing phases of the system's development. Of major importance is that this model be capable of utilizing data concerning the human component's association with or cause of malfunctions.

Purpose of the Model

The model as described above would be a diagnostic tool which would be used continuously during development testing to measure "how well" a system conforms with its design objectives. The need for such a diagnostic tool has become apparent in connection with the development testing of missile systems. These testing programs have been plagued with what has been considered a significant number of malfunction failure reports citing or implying the human component as "the cause" of the malfunction[12]**. Therefore it is important that the relationships of these "human-initiated malfunctions" to other types of malfunctions or failures and to the capability of the system to perform in accordance with the system's design objectives are understood.

*In this report, the term system is often used to mean weapon system and the two terms are used interchangeably where such a meaning is obvious.

**References are listed on page 41.

A Working Hypothesis

In order to place the human-initiated malfunction into its proper perspective, certain pleas are made throughout this report for definitions, standards, specifications, management support, and proper respect for the problem. In addition to these more obvious requirements for placing the problem into a manageable state, it was necessary to establish a working hypothesis to use as a guide during the research period. This hypothesis has neither been proven nor disproven during the research program. The background information leading to the hypothesis follows.

The project has been basically concerned with malfunctions occurring in the research and development test phases of system development. In any engineering development program this stage of the development of a product or system is known as the stage where the bugs are worked out. It is in the development testing stage that the workability of a system is demonstrated. This stage is the transition from research and design engineering to the operational deployment of the system. It is in this development period that unforeseen interactions between components, design oversights, and other unpredictable events occur. During the development testing period the occurrence of these events is not necessarily looked upon in a derogatory sense, but more generally as everyday encounters which must eventually be surmounted before a final product evolves. Many human-initiated malfunctions are of this type, i.e., they must occur and be corrected before workable weapon systems are produced.

The human-initiated malfunction can be hypothesized to be analogous to an equipment component with a short mean life length or a system with a short mean-time between failures. This analogy is believed to be acceptable because many human-initiated malfunctions have to occur and be corrected before any time can be logged on the system under test. Eventually, as the development of a system progresses, most correctable human-initiated malfunctions are eliminated until, for the most part, only a random human error component remains. In other words, most human-initiated malfunctions for which an assignable cause can be found are eliminated. The term assignable cause is used in industrial quality-control activities to describe the causes of variation other than those due to chance. These assignable causes are relatively large variations which are usually attributable to such special causes as differences among machines, among workers, among materials, and differences in their relationships to one another. Equipment failures during development follow a pattern similar to human-initiated malfunctions except that they occur subsequent to the occurrence of human-initiated malfunctions in time. Together these phenomena are often referred to as reliability growth. Many human-initiated malfunctions can only be measured in terms of delays in development testing and are not directly measurable in terms of system performance. A system which matures into a system which meets its design objectives is usually characterized either by random or wearout-type failures. Failures due to other causes have been previously eliminated in development.

The severity of a malfunction as well as the reliability growth aspect of the problem must also be considered. Severity as used herein refers to the degree of performance degradation which the weapon system suffers as a result of the malfunction. Due to random influences, a system seldom achieves ideal characteristics, i.e., some tolerance about its primary performance measure is considered acceptable. This performance may or may not suffer as a result of various types of malfunctions. Trends during development should reveal true system performance, e.g., for a good system there should be a marked differentiation between reliable performance as designed and the performance still attainable after failure of a component. In other words, it is not to be expected that system performance improves if a component fails. In development testing a degree of performance degradation can be associated with each bug revealed and corrected. The sequence of when the more and less severe failures occur is of little importance, but for a properly developing system it would be expected that, on the average, performance would show a gradual improvement. This improvement is not necessarily monotonically increasing because all design changes do not improve system performance. In fact, some design changes which are intended as corrective actions may actually degrade performance so that only an average improvement is observed. Concurrently as more systems enter development testing programs and the scope of the testing increases there is a corresponding increase in the apparent incidence of malfunctions reported. Thus, it may be assumed that, in the development of an acceptable system, the following general trends should be observed as a function of development testing time:

(1) A continual decrease in the ratio of human initiated-to-equipment failures following an early peak.

(2) An average improvement in performance.

(3) An increase in malfunctions as the scope of testing increases which in turn may be reflected as an apparent increase in failure rate.

These trends are taken as a working hypothesis for the research program.

With the above as a working hypothesis, the problem will next be examined in both a missile and premissile context in order to gain a more thorough understanding of the significance of the human-initiated malfunction phenomenon.

The Problem in a Premissile Context

To gain a perspective for this problem, one can conjecture about how this problem was handled with the development of previous premissile weapon systems. The new magnitudes and dimensions introduced by missiles have in many cases injected finer tolerances than were previously required by manned aircraft. When compared with aircraft design, it is obvious that missile design must compensate for the subtle corrections and additions which test pilots knowingly and unknowingly added to the system, thus not only allowing it, but helping the system to achieve its design objectives. Perhaps of more importance is the understanding of how the information concerning human-initiated failures was fed back to the appropriate source for subsequent correction. It may be considered that the test pilot with his considerable influence represented the main component in the feedback loop which initiated orders for correcting situations causing human-initiated malfunctions. Test pilots, both military and contractor, to a great extent established whether a military airplane performed in accordance with its design objectives or not. The test pilots' reports and interpretation of flight test data provided a measuring and feedback mechanism which is not available in the missile industry. The absence of this human interpretation adds to the complexity of the problem faced in missile testing. Also, aircraft flight test procedures with their usual preflight checks no doubt revealed many malfunctions of the variety which would be classified as human initiated today. However, these malfunctions were probably corrected on the spot and never became a part of the official record.

The Problem in a Missile Context

The problem as it exists in the missile industry is in many respects more subtle. In an effort to learn everything possible about weapon system performance and to isolate the cause of every failure, the missile industry has tried to record as many items of information as possible during each test. Data collection systems have been organized to process this information in the hope that critical areas will be automatically pinpointed upon receipt of input data. One of the results of this activity is that many gross statistics which have no practical significance have been compiled, but are nevertheless considered indicators of critical areas. In the rush to correct these critical areas specialized techniques have been developed. The application of special techniques to critical areas indicated by new data collection systems will not correct such problem areas as the human-initiated malfunction overnight. The human-initiated malfunction has been present for ages. Apparently, it becomes more significant with each phase and sophistication of the Industrial Revolution. It has come to the forefront with the latest stage of automation. To reduce it to an acceptable level again, one must examine the organization required to correct it, the time relationships involved in bringing it to the forefront, and the information flow and data processing which cause it to be presented today as a significant statistic.

What Is a Human-Initiated Malfunction?

In weapons systems testing, on-the-spot individual judgment is used in many cases to evaluate the cause of a malfunction. When so judged and reported, this action triggers a chain of events which results in orders for the malfunction to be corrected. There are no established criteria for judging whether a malfunction is human initiated or not. In reviewing malfunction data, one team of researchers used the following procedure to distinguish between human-initiated malfunctions and those that were equipment initiated:

> "Each of the malfunction reports collected was analyzed and the malfunction was classified as either human initiated or equipment initiated. An equipment failure or an unscheduled hold considered to be human initiated if the human component could be clearly identified as the causative agent in the immediate train of events leading to the equipment failure. Equipment failures that could be ascribed to such causes as misassembly, mishandling, or misadjustment by the human operator were identified as human initiated, but failures that ultimately might be ascribed to such causes as misdesign or improper specification were not so identified."(12)

In addition, many of the malfunctions classified as human initiated are more concerned with the support, maintenance, and checkout equipment than with the missile proper. Since these auxiliary subsystems are not always developed concurrently with the primary system, this time differential by itself produces many problems and malfunctions when the primary system and subsystems are integrated. Many of these malfunctions result from improper process specifications, functional test specifications, job descriptions, etc.

Several classifications have been suggested for malfunctions of both equipment type and the human-initiated type. Detailed classifications of equipment-type malfunctions are available and are in use. Adequate detailed classifications of the human-type malfunction are not generally available.

The problem of differentiating between the human initiated and the non-human-initiated malfunction is one of defining a fine line or arbitrary threshold since all malfunctions are caused by human action or decisions (excluding "Acts of God"). This threshold is neither firmly nor consistently defined and is subject to considerable variation as a function of the reporting source. Occasionally, classifications that have been suggested include a general human error category, but supporting documents concerning human factors testing and evaluation supply only vague references and generalities about human-initiated malfunctions. These documents have evolved from human factors handbooks, e.g., Tufts Handbook,[14] and Van Cott and Altman,[15,16] experimental psychological testing schemes, human factors check lists, etc. When compared with human-initiated malfunction classifications, equipment malfunctions classifications appear complete and well refined. Shapero[12] sums the situation as follows:

> "Whereas the specific equipment component is discrete and identifiable by name and numbers, no equivalent tags for the pertinent human actions or operations have been integrated into the failure-reporting system. Only large activity categories, such as, "Inspection", "Test", or "Operations", are provided. These are too gross to permit the identification of specific actions involved in a failure event, and therefore do not aid in the necessary analytical re-creation of the circumstances of failure."

Human-Initiated Malfunction Criteria

The lack of consistent qualitative and quantitative definitions and standards poses a serious handicap for any modeling activities which attempt to use empirical data. Likewise, any inferences drawn from such data are subject to serious criticism. In addition to the need for data collection standards, the biases injected in such data because of testing sequences, learning influences, design changes, etc., deserve considerable attention. Not to be overlooked when establishing standards and data collection facilities is the supporting organization. Unless human factors testing groups are given proper management support, their technical effort will be acting in a vacuum. Management support should be such that the human factors effort is given the authority to design, measure, and predict, using scientific methods, and can be held directly accountable or responsible for the successes and failures of their efforts. The use of standards and specifications, the temporal sequence of testing and data collection, and the management organizations supporting these activities all are important influences upon the development of the model.

This general background provides the setting for the present problem: to develop a model, capable of accepting human-initiated malfunction data, which can measure the agreement between the actual performance characteristics being achieved by the system and the design objectives established for the system. In one sense, the problem can be considered as one of quality-control measurement, i.e., what critical human factors tests or statistics derived therefrom measure the agreement between the actual system performance and the design objectives established for it.

Discussion of the Research Effort

In the exploration of this problem, it was necessary to make certain basic assumptions. These assumptions deal mainly with the conditions under which a malfunction is classified and recorded. As has been pointed out in previous research(8,11,12), the human factors specialist too infrequently commits his recommendations to paper as an engineer does when he makes a drawing or orders hardware by a bill of material. With the aid of engineering drawings, inspection personnel can diagnose, analyze, and explicitly reference equipment failure, incorrect assemblies, improper part applications, and similar malfunctions. Equivalent documents for the detection of human-initiated malfunctions are necessary. Currently any human factors specifications are usually found implicit in other documents whose purpose is to delineate test procedures, fabrication details, and other instructions too complex for inclusion in engineering drawings. If human-initiated malfunctions are to be correctly classified, some additional document which can be properly labled a human factors specification outlining the correct human operations must become a reality. Thus, a necessary assumption in attempting to develop a model to make performance predictions from human-initiated malfunction data is that appropriate human factors specifications or specific classifications exist. This leads to a development of either new human factors specifications or specific classifications of possible human-initiated malfunctions currently found in process and functional test specifications. These specifications must be explicit to such a degree that variations due to interpretation by test inspection personnel are minimized.

A second assumption which follows directly is that standard inspection procedures exist or can be developed. This assumption along with the first eliminates or tends to eliminate subjective biases introduced into the malfunction reporting system. With human factors specifications or human factors checkpoints included in other specifications, it becomes possible for inspectors to make explicit checks. This would provide a means for qualifying human-initiated malfunctions; it would also provide a consistent classification method in place of the current ex post facto subjective interpretation of previously collected failure data. These two assumptions then form the basis for the development of a human factors testing program oriented toward the reduction and control of the inherent human-initiated malfunctions found in research and development testing phases. A further need still exists in the management domain – that of specifying a means for accomplishing a human factors testing program.

Recent Research Bearing on the Problem

Several research programs in the recent past have a bearing on this problem. For the most part these efforts attempt to delineate the problem area as seen from several viewpoints and thus quite different approaches to the problem were made. A few of the programs which have had significant influence upon this work are discussed below.

One of the earlier of these works is one by Shapero, Cooper, et al.,[12] whose efforts gave rise to the work reported herein. These researchers coined, and most appropriately so, the term human-initiated malfunction. From their limited survey of nine missile systems research and development programs in which they attempted to ascertain the status of human factors testing, they reported

> "Little, if any, formal or systematic human factors engineering testing is being conducted...Present malfunction reporting systems are inadequate in identifying human-initiated malfunctions; or for providing analyzable data pertinent to human-initiated malfunctions..."

In their analysis of 4248 malfunction reports it was concluded that a significant portion were of the human-initiated type. Of the nine systems investigated, the portion of human-initiated malfunctions varied from 20 to 53 per cent of the reported equipment failures and 15 to 23 per cent of the unscheduled holds. As the authors stated, their classification of human-initiated malfunctions may contain considerable subjective bias because the reporting procedures in use do not properly reflect failures of a human-initiated nature. The researchers further state that – "The data currently being collected concerning the circumstances and characteristics of an individual malfunction vary somewhat from reporting system to reporting system".

These researchers also present an argument supporting the development of standardized human factors techniques. In research experimentation and engineering testing a degree of accuracy is required, which is not yet found in the equally important human factors testing area. This situation exists partly because of the lack of requirements set forth by the organization charged with system development, and partly because there are no widely accepted criteria regarding human factors testing. These researchers supplement their argument with a proposal for human factors test programs. However, their proposal is one of concept and not of action or organization.

Story[13], in her review of evaluation techniques currently in use, discusses the determination of adequate man-machine system performance criteria. Four categories suggested, material, mechanistic, logical, and final criteria, correspond roughly to the "building block" testing program commonly used. By "building block" is meant a program of testing from the component material level through building of operational units to testing the total system in either a simulated or an operational environment or both. From the points of view of the designer, builder, tester and controller of man-machine systems, determination and application of adequate criteria should be made as soon as it is feasible to test any aspect of system performance. This supports the contention that human factors methods should be introduced as early as possible in a weapon system development and that criteria establishing what constitutes a human-initiated malfunction be determined and utilized regardless of where the malfunction appears. Recognition that a human factors development testing program is as important as, and is integral with, the development of hardware subsystems from a systems engineering point of view has not yet been fully realized.

Meister[7] states that a system encompasses not only groups of machine subsystems performing "... a unified operational function but also the personnel who operate and maintain these machines". These system functions not only deal "... with the ultimate tactical operation – but also with all the supportive operations: checkout, preventative maintenance, logistics, planning, etc...." and all involve varied personnel functions. Therefore, when talking about system reliability one must not only be concerned with hardware reliability, but also with personnel reliability – efficiency of the personnel who operate and maintain the hardware systems.

Meister also discusses one point often overlooked, operational system reliability. Although this topic is not the concern of this report, it has bearing on the problem at hand. Often reliability drops after a system is transferred to the user. Usually the hardware component of reliability remains unchanged; it is the human component that causes the resultant drop. Various answers are given for this drop in system reliability. Foremost among these is that the operational personnel "... are not initially as adequately trained as our civilian technicians; hence the former encounter greater difficulties". But more important is the fact that exactly the same thing happens when a system is transferred from research and development personnel to field testing personnel and from one group to another within research and development personnel. Why? Perhaps the most appropriate answer lies in the fact that generally no centralized management control exists which can coordinate human factors design and testing activity at all levels of research and development. Kucij[6] and Drose[3] have both made pleas for better failure data classification and evaluation in terms of a weapon system's objectives.

A consistent theme underlying all this previous research is that human engineering testing programs should be developed as early as possible in system development and the importance of this function should be more fully recognized.

CHAPTER 2

RELIABILITY GROWTH AND DEVELOPMENT TESTING

Brief

During development testing the satisfactory performance of a weapon system should be established or the system shown to be unsatisfactory. Included within the concept of satisfactory performance is the idea of reliable performance. However, since reliability is usually measured in statistical terms, the transitory nature of the development testing period makes it difficult to establish the operating reliability of weapon systems. This section of the report deals with some of the reliability aspects of human-initiated malfunctions which occur in development testing.

Reliability Definitions and Component Life Length

Reliability Definitions

As a secondary objective, the Statement of Work for the research program states that the model developed should be useful for predicting system reliability. Reliability is empirically appraised by statistical analysis of failure data. Since human-initiated malfunction data are a part of the failure data, they are implicitly included in reliability analyses. It is therefore necessary to review the accepted definitions and formulation of reliability from an equipment viewpoint. Definitions of reliability are not consistent throughout the weapon system industry, but for purposes of this program the more commonly accepted definition of equipment reliability will be used. This definition is: the reliability of a weapon system is the probability that the system will give satisfactory performance for a given period of time when used in the manner and for the purpose intended. Such a definition at first consideration seems to be quite stringent. Further reflection shows that considerable latitude and interpretation can be given particularly to such questions as:

(1) Within what tolerance limits does satisfactory performance lie?

(2) How is the given time period established?

(3) For a weapon system, how are all the possible uses going to be predicted and defined?

(4) How closely is the intended purpose actually controlled?

It is to be parenthetically noted that in the past there have been many bomber airplanes converted to tankers and reconnaissance vehicles, fighters to bombers, etc., many with regrettable consequences. In other words, unsatisfactory performance and unreliable operation resulted from violation of the idea – when used in the manner and for the purpose intended.

Because of the usual looseness in defining satisfactory system performance, several difficulties arise in trying to relate component part reliability to subsystem reliability and, in turn, subsystem to system reliability. Unsatisfactory assumptions are often made regarding the independency of the interactions of failed and unfailed components. More about the relationship between reliability and performance will follow, but it is safe to say, in general, there currently is no neat, unified, and workable method for quantifying reliability. Anyone selecting a set of rules for use in reliability evaluation is simultaneously making himself a target for future censure. It is believed (and hoped) that a certain maturity has permeated the workers in the reliability field so that it is now realized that there are no universal definitions for reliability nor any over-all panacea for measuring reliability.

Life-Length Approach

At the component part level, reliability is usually approached by life-length or time-to-failure measurement. There are several ways of presenting life-length or failure models. However, failure processes are usually modeled mathematically by assuming that life lengths of nominally identical objects subjected to a given time-stress pattern have a probability distribution. Such a model assumes that there exists a time-dependent probability, $F(t)$, such that an object selected randomly from a population will fail during the time interval, 0 to t. This object sustains damage or wear in time described by a nonnegative function, $\Delta(t)$, such that $\Delta(t)$ is the instantaneous damage to the object at time, $t \geq t_a$, where t_a is a minimum life value. If the component fails at time, t, we call t the life length under the damage function, Δ. It is assumed that the probability of life length will depend on the damage function, Δ. Indicating this dependence, we have

$$P[T \leq t \mid \Delta] = F(t; \Delta) \text{ for all } t > t_a, \qquad (2\text{-}1)$$

and

$$d/dt \; F(t; \Delta) = f(t; \Delta) \qquad t > t_a, \qquad (2\text{-}2)$$

where T is the observed life length.

The hazard rate, z, is defined as the instantaneous failure rate at time, t, conditional upon the nonfailure of the object prior to the time, t, or

$$z(t) = \frac{f(t; \Delta)}{1 - F(t; \Delta)} \quad . \tag{2-3}$$

In other words, the hazard rate is the failure rate of an object at a certain instant given that the object has survived until that instant. It is interesting to note that this formulation is the same as the "force of mortality" used by actuaries in mortality statistics and the "intensity"[5] function used in other branches of statistics. The hazard rate is the salient parameter which abstractly relates probability distributions to the underlying failure mechanism. It is the ordinate of the probability distribution divided by the area under the distribution which is greater than t. Integration of Equation (2-3) gives

$$F(t; \Delta) = 1 - \exp\left[- \int_{t_a}^{t} z(s) \, ds \right] \tag{2-4}$$

Reliability is taken as the probability of no failure, that is,

$$R(t; \Delta) = 1 - F(t; \Delta) = \exp\left[- \int_{t_a}^{t} z(s) ds \right] \tag{2-5}$$

where R = P [no component failure up to time t].

If the probability distribution of the time to failure is known, the hazard rate can be determined. However, it is in determining the probability distribution where many practical problems in reliability are concentrated. A constant hazard rate implies that old, nonfailed items are as good as new items given only failed or unfailed conditions. If the hazard rate increases with time, it is usually inferred that a part is aging or being progressively damaged. If the hazard rate decreases with time, the implication may be that a group of old items is better than a group of new items, or that items improve with age. In such a situation it may be a considered advantageous to shake down items prior to use in order to improve reliability. This phenomenon, early failure, is often referred to as infant mortality.

Table 1 presents the characteristics of the hazard rate for five commonly used failure density functions. Plots of these hazard rates are shown in Figure 1.

TABLE 1. CHARACTERISTICS OF THE HAZARD FUNCTION z(t) FOR FIVE FAILURE DENSITY FUNCTIONS

	Probability Density Function	Changes in z(t) as a Function of Time	$\lim_{t \to \infty} z(t)$
(1) Lognormal	$\frac{1}{t \sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\ln t - \mu)^2}$	Increases until some time $t > e^{\mu - \sigma^2}$ and then decreases	0
(2) Exponential	$\frac{1}{\tau} e^{-t/\tau}$	Constant	$\frac{1}{\tau}$
(3) Gamma	$\frac{1}{\tau \Gamma(r)} \left(\frac{t}{\tau}\right)^{r-1} e^{-t/\tau}$	Increases monotonically with time to a limiting value (r > 1).	$\frac{1}{\tau}$
(4) Normal	$\frac{1}{\sqrt{2 \pi} \sigma} e^{-\frac{1}{2\sigma^2}(t - \mu)^2}$	Increases monotonically without limit	∞
(5) Weibull	$\left(\frac{c}{r}\right)\left(\frac{t}{r}\right)^{c-1} e^{-(t/r)^c}$	Increases monotonically without limit (c > 1)	∞

Hazard Rate and System Life

Figure 2 is a sketch of a hazard rate for the life of a typical system. It should be noted that Figure 2 includes the development stage while many similar representations start with the operational life period. Early in the development phase the hazard rate is low and slowly increases. Failures occurring at this stage could be referred to as prenatal. At the time when a system becomes operational the hazard rate is near its maximum. Often development testing continues after certain earlier produced units have become operational. It therefore may be a matter of opinion when it comes to establishing the exact start of the operational phase. Following the maximum the slope changes and the hazard rate decreases to a constant value. The beginning of a weapon systems life is usually characterized by this infant mortality period. Some observers would rather consider this burn-in or shake-down period prior to the actual beginning of operational life. The transition period is often so hazy that any arbitrary differentiation would suffice for most purposes. Subsequent to the constant hazard rate period comes the wearout stage wherein the hazard rate again increases and may reach another peak and decrease. This final stage for a weapons system could be analogous to a Model T automobile in highway service today.

Not all systems meet the general characteristics presented in Figure 2. For instance if the increase representing wearout is shifted to the left such that the wearout period immediately follows the infant mortality, then it may be concluded that the equipment is only partially developed. In general, for a good weapon system it is desirable to have eliminated all early failures as soon as possible and to have wearouts occur long after the designed operational life of the system. In other words the negative slope representing early failures should be steep and the wearout area should be distinctly separated from early failures. Wearout should not occur before the system is debugged.

Little is known about hazard rates early in the development phase, probably because of the complete fluidity of the situation. In most weapon systems development programs several systems are produced for test. However, not all the systems are subjected to the same test. For example, some may be subjected to structural tests, some to environmental tests, and others to functional tests. These systems are usually constructed sequentially and therefore the number of systems under test increases as the development testing program proceeds. In addition to the changing number, tests are often censored because systems under test are destroyed. As a result there is little control in a statistical sense. Design changes are the rule rather than the exception. Several learning phenomena are occurring and as a result each successive system may differ considerably from the previous. Another factor to be included is the personnel training problems, both in the training of test personnel and of future operating personnel. It is virtually impossible to approach controlled conditions from

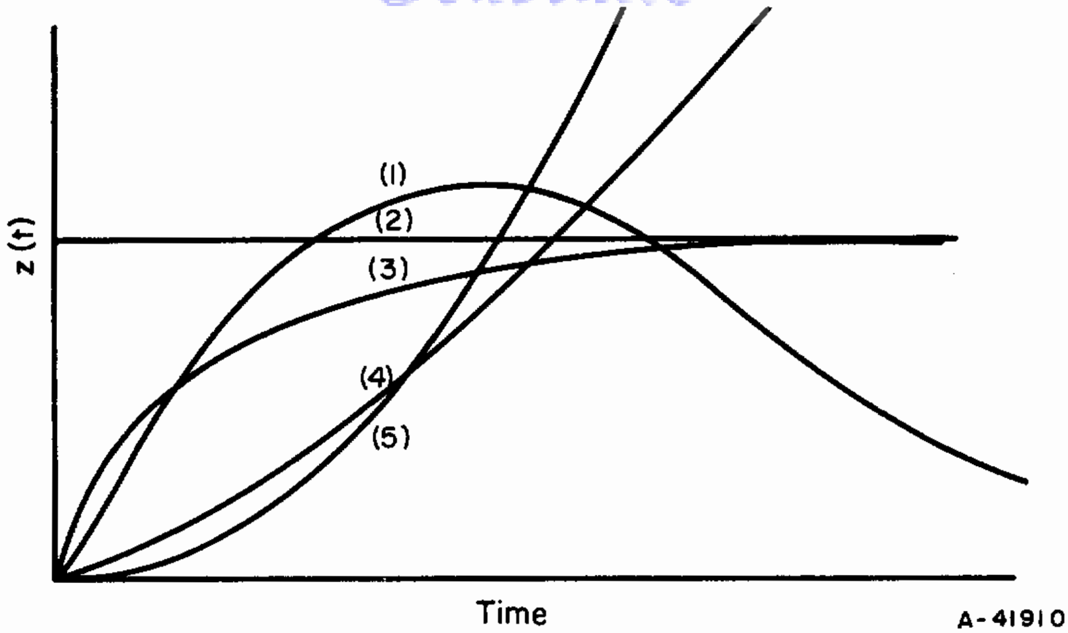

FIGURE 1. HAZARD FUNCTIONS FOR SEVERAL FAILURE DISTRIBUTION FUNCTIONS

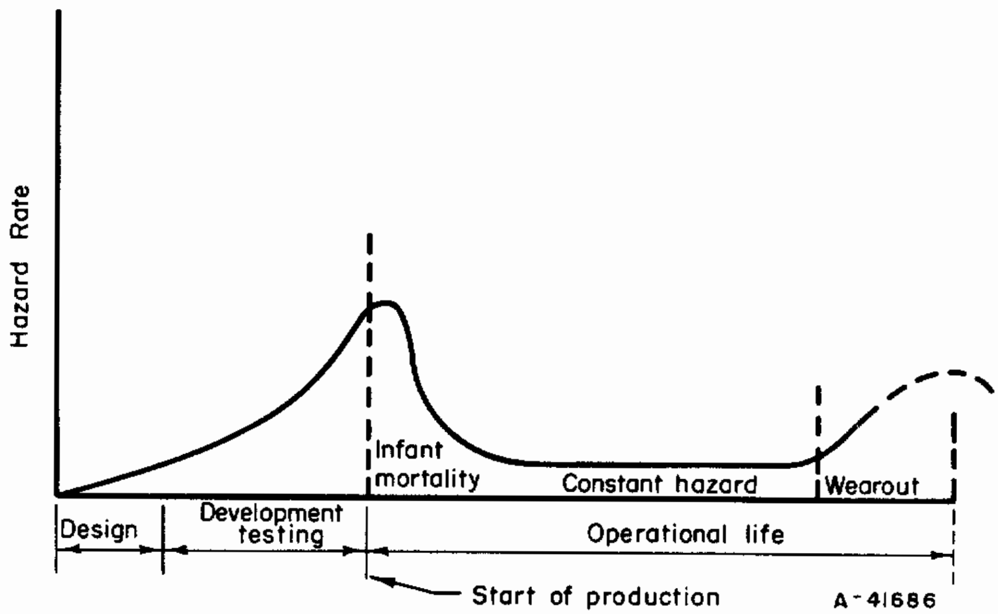

FIGURE 2. MORTALITY CURVE SHOWING TYPES OF FAILURE CONSIDERED IN WEAPON SYSTEMS RELIABILITY

test to test. For all of the above reasons it is considered difficult to determine the desired hazard rate profile during development testing.

If the hazard rate profile of an actual weapon system is as depicted in Figure 2, then an increase in failures and consequently a decrease in reliability can be expected during development testing. Part of the increase in the hazard rate during this testing phase is merely a reflection of the increase in the number of systems under test. However, for a goodly portion of the development testing phase the apparent reliability of a system decreases. What is disconcerting about this is that, although the development of a system may be progressing satisfactorily, the apparent reliability as shown by malfunction data may be steadily decreasing as the operational phase is approached. Figure 3 is a graph of actual malfunction data for two missile systems collected during missile development which shows this increase in the number of malfunctions.

It seems quite likely that this decrease in reliability has little to do with the eventual inherent reliability of the fully developed, satisfactory weapon system. If all such malfunctions which were assignable to development test procedures, oversights due to the newness of the system, unfamiliarity, learning new techniques, etc., were separated out, then the hazard rate representing the remaining malfunctions would approach the same hazard rate which characterizes the operational life of a good system. However, for the many reasons stated previously about the transitory conditions during development testing such data could be deceptive and undependable, e.g., when treated on the individual component level, sample sizes would be too small for meaningful statistical tests to be made.

For purposes of analysis then it is expedient to consider human-initiated and other malfunctions closely associated with development testing as being analogous to component parts with very short mean lives. As such, these are possible defects which must be screened out by testing. When considering some of the various types of human-initiated malfunctions, the similarity in time of occurrence of these and component failures with short mean lives is obvious. The following are examples of human-initiated malfunctions which can occur before any test time is accumulated on the system under test or shortly thereafter:

(1) Malfunctions due to following incorrect or incomplete test specifications

(2) Malfunctions due to not following instructions

(3) Misassembly or installation error

(4) Improper use of tools, test equipment, or servicing of equipment

(5) Inspection error (either safety or quality control).

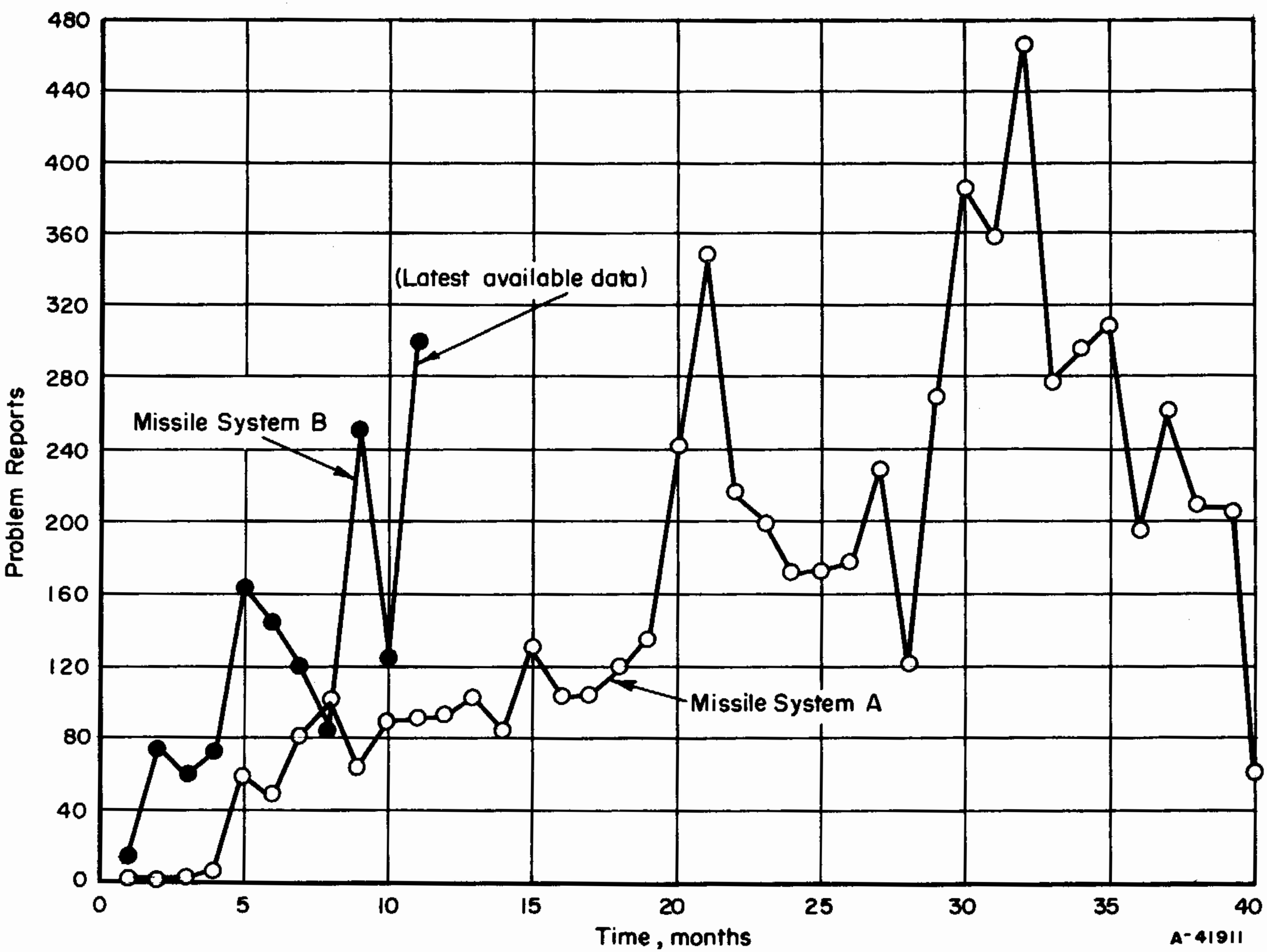

FIGURE 3. PROBLEM REPORTS BY MONTH FOR TWO WEAPON SYSTEMS

Under this analogy, human-initiated malfunctions are of the early failure type and the majority of these are removed early in development. Data from a missile systems development program as shown in Figure 4 (although inconclusive) tend to indicate this trend. To smooth these curves, fifth degree polynominals were fit to the data by the method of least squares utilizing orthogonal polynominals (see Figure 5).

A test program for a weapon system developing satisfactorily is perhaps characterized by an initial rash of human-initiated malfunctions which decrease as the test program progresses. Percentagewise these malfunctions may reach significant proportions. Eventually, as all human-initiated malfunctions for which assignable causes can be found are corrected, the human error (random error) will be reduced to acceptable tolerances.

For obvious reasons it is more desirable to talk of reliability growth instead of decline. However, if reliability can decrease during development testing as previously conjectured, then negative growths may have to be considered. There are some idealized models available for measuring reliability growth. How practical these models are for utilizing field data is not known. H. K. Weiss(17) has presented a model for estimating reliability growth in a complex system. This model was developed to consider the case in which the reliability of a system is increased by the process of testing the system until failure, identifying the cause of the failure, correcting the cause, and retesting. This sequence is repeated until a satisfactory level of reliability is achieved. The proposition put forth here seems to be more applicable to the infant mortality portion of the hazard rate curve although the model can be adapted to represent areas where there is a negative reliability growth. Weiss treats the case where the system is subject to Poisson-type failure and gives examples for constant mean time to failure and for mean time to failure changes by constant percentage on each test. Formulas are developed for the cases where equal and unequal failure rates are present in the system. Intuitively the latter case is the more likely to occur. Examples are also presented to show that if an infinite number of possible failure sources are present, the testing procedures produce a steady growth in mean time to failure since the sources of the high failure rates are most likely to be eliminated early in the testing process. Sources with very low failure rates may never be eliminated in the operational life of the system. Conversely an error source which has a very low mean time to failure is almost certain to be removed on first trial.

Other growth curve techniques, such as the "logistic", Gompertz, etc., are available to describe the reliability growth phenomena. It may be more practical to fit the growth curve to the hazard rate for the development period and then perform the integration indicated in Equation (2-5) to obtain the reliability function. Because observations from malfunctions occurring during a development testing program are usually not stochastically

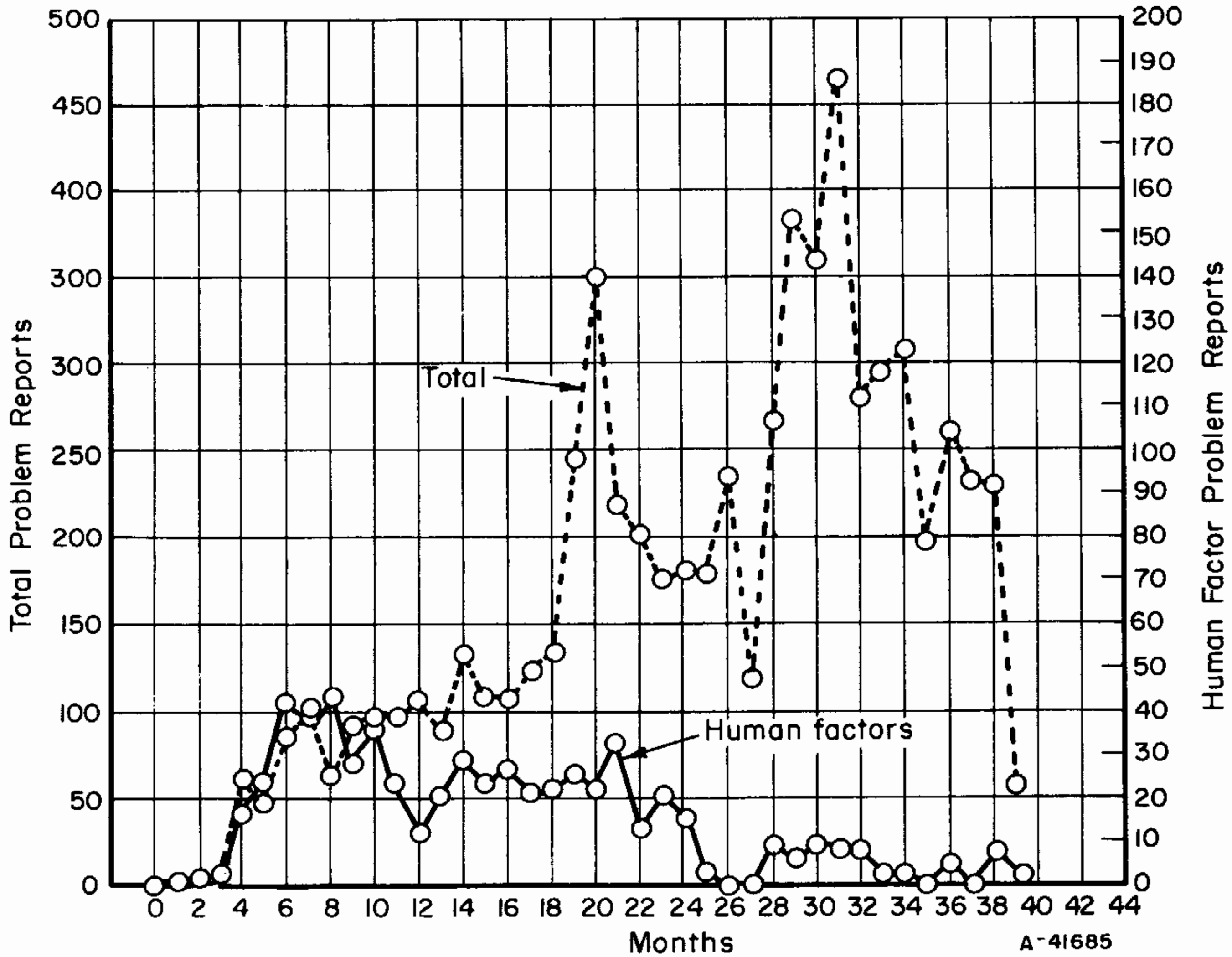

FIGURE 4. HUMAN FACTORS AND TOTAL PROBLEM REPORTS BY MONTH

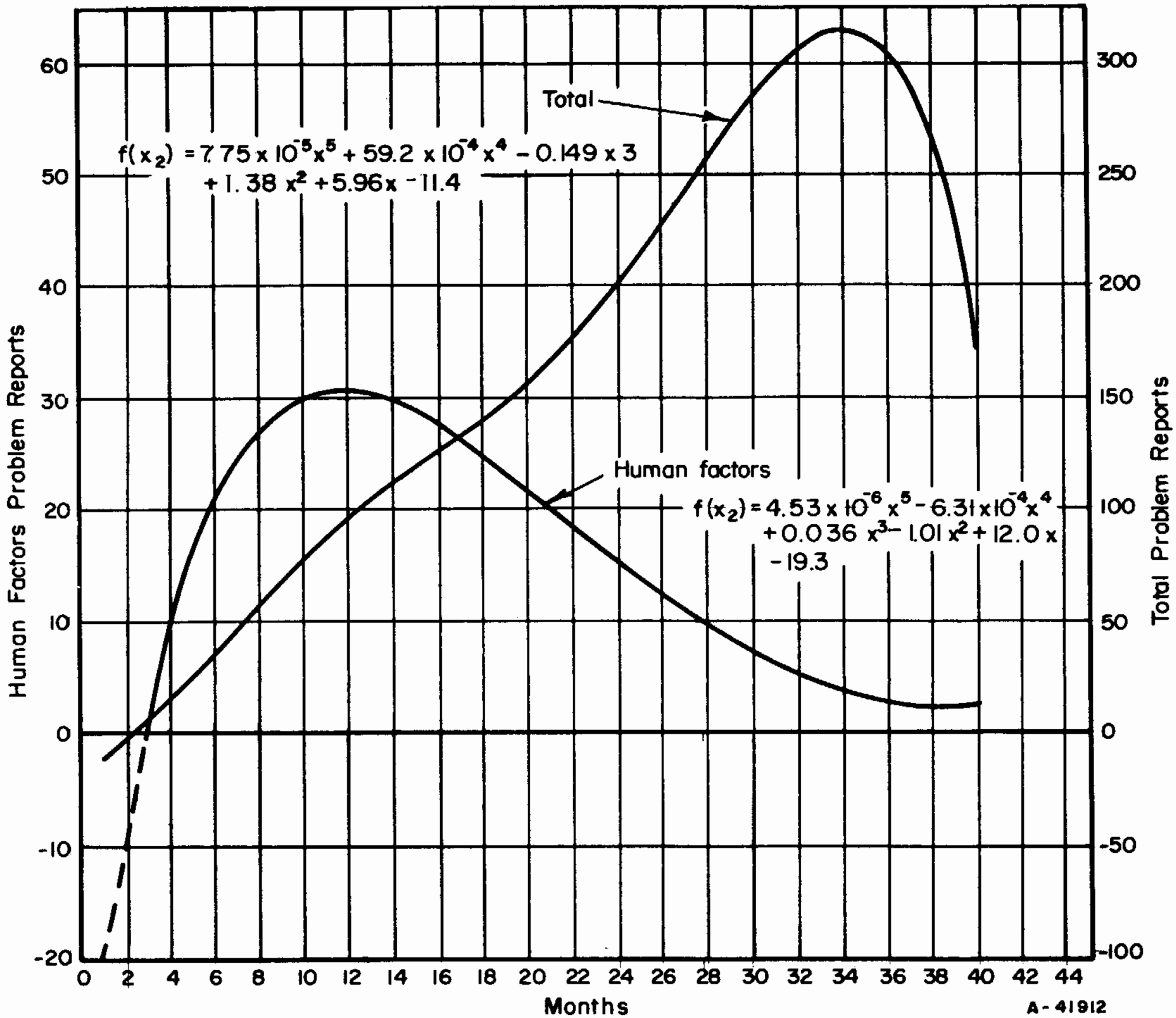

FIGURE 5. PROBLEM REPORTS BY MONTH, SMOOTHED CURVES

independent, it is recommended that growth curve techniques be applied with the idea of being descriptive only and not in hopes of theoretically describing any basic growth process. Boyle[2] has presented a sequential analysis method for using test data to determine whether a missile is achieving the reliability growth as predicted by a predetermined growth curve.

The reliability as measured in the development testing phase is not particularly useful for predicting the future reliability of a weapon system. If a growth curve is fitted to a hazard rate representing the development testing stage only, and if it is shown to approach a limit, then this limit can be used to determine at what level the operational reliability would start [see Equation (2-5)]. This level would be considerably more realistic than some overestimated reliability figures used in the past. Should it be possible to separate the hazard rate into a monotone function for the development period, some properties of the hazard rate presented by Barlow and Marshall[1] could prove useful. On the basis that it is very difficult to distinguish between various parametric families of failure distributions from empirical data, Barlow and Marshall suggest that it is desirable to confine attention to a nonparametric class of distributions determined by properties of the hazard rate. These authors investigate various bounds on survival probabilities (reliability), number of renewals, and other allied properties.

Conclusions

It appears that the best reliability prediction that can be made from development testing is the initial reliability at the outset of operational life. This reliability is likely to be low due to the infant mortality phenomena when operational life begins. Human-initiated malfunctions are believed to occur primarily in the early portions of development testing. The human-initiated malfunctions reach an early peak and decrease as development continues, but equipment failures continue to increase until operational life of the system begins. Upon the completion of development testing, human-initiated malfunctions should be reduced to an acceptable error rate and should be completely random in nature. Other than monitoring these general trends it is doubtful whether any controls can be put upon reliability growth functions during the development period. Analysis can be done only on specific systems and individual judgement of these analysts intimately connected with the system must be relied upon in each case. In other words there is no over-all panacea or general solution; each system represents an individual case.

CHAPTER 3

PERFORMANCE EVALUATION DURING DEVELOPMENT TESTING

Brief

A method is presented for incorporating reliability and system performance in one system measure. On the basis of this measure, it is shown how system performance may be monitored during development testing.

Performance Measures

The idea of system performance carries with it the connotation of equipment or man-machine systems, etc., operating in a perfunctory manner and accomplishing their intended function. Any deviations from this normal operating condition due to component failure, breakage, or running out of tolerance is usually associated with reliability. The interrelations of system performance and reliability often tend to obscure the measurement and prediction of either or both. In Chapter 2 it was noted that a weakness in the usual definition of reliability is found in setting the limits on what is considered to be satisfactory performance of a system. It is often observed that the failure of a single peripheral component may not perceptibly degrade the performance of a system and conversely the failure of any one of a few "vital" components may cause the complete loss of system performance. Somewhere between complete system failure and specification performance there are intermediate degrees of system performance which can be related to component failures or partial subsystem failure. In this intermediate state a judgement is required concerning how much performance degradation is acceptable before the system is considered to have failed. In making this judgement two problems are encountered: first, that of relating individual component reliability to system reliability, and second, that of determining what minimal level of system performance constitutes acceptable system performance.

Reliability and System Performance

Several schemes have been proposed to relate component and system reliability. To consider that any one component failure is equivalent to failure of the entire system is to overgeneralize. Another scheme is to divide components into two groups: one group that causes "significant failures" which degrade system performance severely, and the other group that causes "insignificant failures" which can be ignored in computing system reliability. This segregation of significant and insignificant failures can be extended further by representing the system as a set of series and parallel (redundant) elements and considering system failure to be equal to the failure of a series element. In such a representation these elements may be individual components, groups of components, or subsystems depending upon their relationship to over-all system performance.

Roberson(10) has proposed a model which alleviates many of the problems mentioned above and attempts to unify reliability and system performance measurement. In this model it is suggested that accuracy be used as the value measure which represents system performance although the model may be extended to many other measures. Accuracy is a good selection as a value measure because it is often the single characteristic which measures the degree to which a system accomplishes its intended function. Such measures as kill and damage probabilities are more inclusive and could be used, but they are not as tractable for development work. Accuracy holds special meaning for this research program since it is customarily associated with bombing systems, missile systems, fire control systems, certain reconnaissance systems, etc., which include a majority of the current weapon systems. The performance of a system may then be characterized by the probability statement

$$P\left\{\theta < \theta_p\right\} = p, \tag{3-1}$$

where θ is the system error (e.g., miss distance) and θ_p is the acceptable system error in achieving probability level of performance p.

Some discussion may be parenthetically inserted here for the use of those not familiar with statistical theory. The miss distance of a bomb or missile from a target may be used as an illustrative example to provide background for the meanings implicit in Equation (3-1). If, in a series of tests (or military operations) it is observed that θ is the coordinate of a burst point of a bomb, then the statement that the population distribution function (or probability density function) for θ is $g(\theta)$ means "if a test is carried out, the probability is $g(\theta)\,d\theta$ that θ will lie between θ and $\theta + d\theta$". For a continuous variate the probability that θ will fall in any interval $a < \theta < b$ is given by the integral

$$P(a < \theta < b) = \int_a^b g(\theta)\,d\theta. \tag{3-2}$$

If the miss distance were to be measured from the target (ground zero), then a would equal zero and Equation (3-2) becomes

$$P(0 < \theta < b) = P(\theta < b) = \int_0^b g(\theta)\, d\theta. \qquad (3\text{-}3)$$

According to its definition, the probability density function must be nonnegative over the defined range of the variate and its integral over the entire range must be unity. Therefore, if it is desired to pick arbitrarily some percentage, b/100, there is some miss distance, θ_b, related to this percentage by $g(\theta)$ and for b/100 per cent of the time the tests are conducted the miss distances will be less than θ_b. In Equation (3-1) this percentage has been called, p, the probability of this level of performance.

Based on Equation (3-1), Roberson* has suggested the following functional relation for reliability and performance:

$$p = R\, A(\theta_p) + (1-R)\, F(\theta_p), \qquad (3\text{-}4)$$

where

R = reliability, i.e., probability of no component failure or no human-initiated malfunction

$A(\theta_p)$ = inherent system accuracy, i.e., probability that $\theta \leq \theta_p$ given no component failure or human-initiated malfunction

$F(\theta_p)$ = failed accuracy, i.e., probability that $\theta \leq \theta_p$, given one or more component failures or human-initiated malfunctions.

The solution θ_p of Equation (3-4) measures the performance of the system and is a function of the value p. Both $A(\theta)$ and $F(\theta)$ are cumulative distribution functions which increase from 0 to one as θ varies from zero to infinity. Typical forms for these functions are shown in Figures 6 and 7. Equation (3-4) resolves many of the previously mentioned conflicts between system reliability and performance. Reliability as defined in Equation (3-4) means the probability of no failure of any component. When used with the $A(\theta)$ and $F(\theta)$ distributions the definition of system reliability based upon individual component reliability becomes more useful. If the distribution, $F(\theta)$, can be obtained, then the problems in setting tolerance limits or satisfactory performance are eliminated, since Equation (3-4) reflects the degree of performance degradation and not complete system failure upon failure of individual components.

*Parts of Mr. Roberson's paper have been reproduced here with the permission of The Journal of the Franklin Institute.

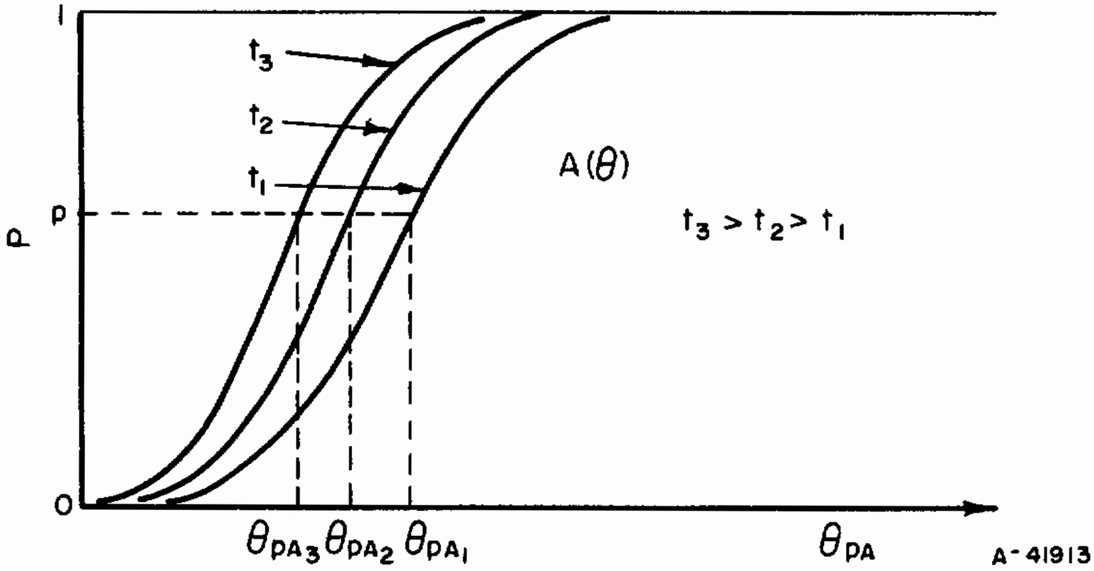

FIGURE 6. TYPICAL CURVES OF A(θ) FOR DIFFERENT DEVELOPMENT TIMES, t_1, t_2, AND t_3

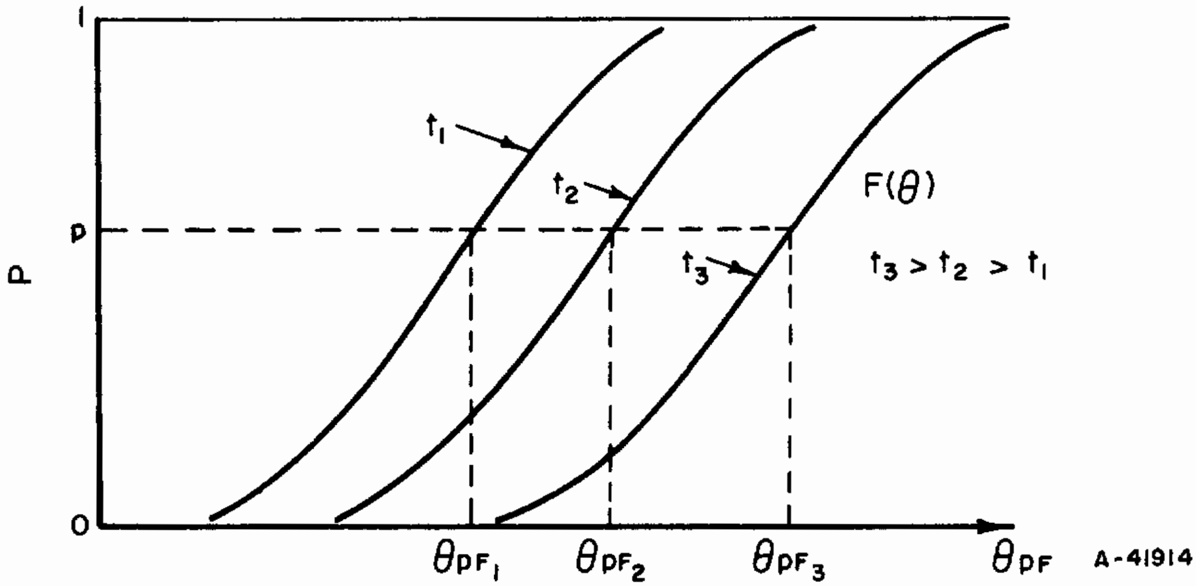

FIGURE 7. TYPICAL CURVES OF F(θ) FOR DIFFERENT DEVELOPMENT TIMES, t_1, t_2, AND t_3

Several interesting conclusions can be drawn from Equation (3-4) and a few are presented below. Three quantities used in presenting these conclusions are defined by Equation (3-5):

$$A(\theta_{pA}) = p$$

$$A(\theta^*) = p/R \quad p < R, \tag{3-5}$$

and

$$F(\theta_{pF}) = p.$$

The quantities θ_{pA} and θ_{pF} and their relationship to the distribution function is shown by dotted lines in Figures 6 and 7, respectively.

The values of θ_p satisfying Equation (3-4) lie between θ_{pA} and θ_{pF} regardless of whichever is the larger. In this way R may be thought of as a weighting function for intermediate points. Considering the case when $\theta_{pF} < \theta_{pA}$, then $\theta_p > \theta_{pA}$ of the system is better when failures occur than when they do not, indicating that $R = 0$ would be an optimum reliability. Obviously, the converse should be a requirement for a properly designed system. Therefore the condition that $\theta_{pF} > \theta_{pA}$ should be used as a criterion for evaluating whether a system has met its design objectives. For $p \geq 1/2$ and $R > p$ then it also can be shown that

$$\theta_{pA} \leq \theta_p \leq \theta^*, \tag{3-6}$$

giving bounds for a particular performance/reliability conditicn. It may also be noted that for $R > p$ the difference between these bounds approaches 0 as R approaches 1.

The distribution $A(\theta)$ represents the inherent accuracy of a system. During development testing the $A(\theta)$ distribution could be empirically determined from the tests which were completed without failure. When measured at successive intervals of time during development it would be expected that the $A(\theta)$ distribution would move to the left as shown in Figure 6 if the system's performance gradually improved during development. As seen in the figure, the error is progressively smaller, i.e., $\theta_{pA_1} > \theta_{pA_2} > \theta_{pA_3}$.

At the completion of a successful development period the $A(\theta)$ distribution should correspond with the designed accuracy of the system. Thus a measure of how well a system conforms to its design objectives can be obtained by comparing the experimentally obtained $A(\theta)$ distribution with the performance accuracy as defined by the design specification. The usual statistical techniques such as "goodness of fit" and "significance" tests could be used for this purpose.

In a similar manner the $F(\theta)$ distribution should tend to move to the right as development progresses indicating a larger decrease in system accuracy upon failure (see Figure 7). In this case it should not be emphasized that the error is becoming larger (which it is), but that the separation between inherent and failed accuracy is becoming larger. If such a change is accompanied by a reliability "growth" the larger error will occur less frequently. The movement of the distribution could be just a change in the shape of the distribution such that it becomes more skewed to the left. The movement of the distribution is based on the idea that a properly designed system will not achieve the same performance in both failed and unfailed states. A completely redundant system could approach this dual state but is not considered here. The limit for this movement would be infinite error upon each failure but this limit has been previously discounted. Therefore, some measure short of this limit should be established. Although not included in performance specifications today, specifications of the future may establish some control on the final $F(\theta)$ distribution for a completely developed weapon system. It is assumed here that initially the $F(\theta)$ distribution will contain many failures which do not degrade system performance. The human-initiated malfunction which does not degrade performance but does delay testing would be a typical example. When these types of malfunctions are eliminated as development progresses it would be intuitively expected that the $F(\theta)$ distribution would contain a larger proportion of more serious malfunctions in terms of system performance. If more is known about the characteristics of either the $A(\theta)$ or $F(\theta)$ distributions, many additional statistical techniques become available. For example, the successive difference between location parameters of the $A(\theta)$ and $F(\theta)$ distributions could be checked as an indication of development progresses.

Severity of Failure

Roberson has also investigated some of the possible relationships between the $A(\theta)$ and $F(\theta)$ distributions. Some of these relationships have interpretations for this research project. Implicitly the $A(\theta)$ and $F(\theta)$ distributions contain a number of parameters from the individual distributions from which they are compounded. Parameters of the $A(\theta)$ distribution may be designated as "accuracy" parameters and parameters of the $F(\theta)$ distribution as "failure-severity" parameters. In certain distributions it is possible to develop relationships between the two to obtain criteria for properly designed systems. For instance, the criterion that the accuracy should not be statistically less in a failed than in an unfailed system puts certain bounds on the relationships between accuracy and failure severity parameters.

In order to exemplify the possible uses of this formulation it is useful to consider a specific case where both the $A(\theta)$ and $F(\theta)$ distributions are circular normal as in Equations (3-7) and (3-8):

$$A(\theta) = 1 - \exp -\theta^2/\alpha^2 \tag{3-7}$$

$$F(\theta) = 1 - \exp -\theta^2/\phi^2. \tag{3-8}$$

α and ϕ are respectively the failed accuracy and failure-severity parameters of the $A(\theta)$ and $F(\theta)$ distributions. Using these distributions, a parameter ξ^* will be derived which indicates the relationship between reliability, failure severity, and system performance. Substituting the circular normal distributions for $A(\theta)$ and $F(\theta)$ in Equation (3-4) gives, after suitable manipulations

$$R \exp(-\theta_p^2/\alpha^2) + (1-R) \exp(-\theta_p^2/\phi^2) = 1 - p. \tag{3-9}$$

Then define σ as another "failure severity" parameter where

$$\sigma = (\phi-\alpha)/\alpha. \tag{3-10}$$

Recalling that in a properly designed system that the accuracy should not be statistically less than in a failed system, it is then necessary that $\phi \geq \alpha$ or $\sigma \geq 0$. Or, failures which do not broaden the distribution have zero severity. If $\sigma = 0$, then the error has the value

$$\theta_{pA} = \alpha\sqrt{\ln(1/1-p)}. \tag{3-11}$$

In Equation (3-11) the symbol θ_{pA} is used since the same accuracy is obtained with $R = 1$ as when $\sigma = 0$. At infinite failure severity, $\sigma = \infty$ then

$$\theta^* = \alpha\sqrt{\ln(R/R-p)},\ (R > p). \tag{3-12}$$

Equation (3-12) gives the same result as if all component failures resulted in complete system failure.

It is convenient to normalize θ_p to θ_{pA} by defining

$$\xi = \theta_p/\theta_{pA}. \tag{3-13}$$

Equation (3-9) may then be rewritten as

$$R\,(1-P)^{-1} + \xi^2 + (1-R)\,(1-p)^{-1} + \xi^2/(1+\sigma)^2 = 1. \tag{3-14}$$

No system degradation due to failure corresponds to $\xi = 1$ and infinite degradation to $\xi = \infty$. Although Equation (3-14) generally cannot be solved explicitly for ξ it is easy to find σ as an explicit function of ξ and R as shown in Equation (3-15):

$$\sigma = -1 + \frac{\xi}{1 + \ln\left[\frac{1 - R(1-p)^{-1} + \xi^2}{1-R}\right] \Big/ \ln(1-p)}, \qquad (3\text{-}15)$$

for $(1 \leq \xi < \infty,\ 0 \leq R \leq p)$,

and $(0 \leq \xi < \xi^*,\ p < R \leq 1)$,

where $\xi^* = \sqrt{\ln(R/R-p)/\ln(1/1-p)}$. (3-16)

Values of ξ^* are important because for values of R in the interval $p < R \leq 1$ these values are asymptotes beyond which ξ cannot increase and these asymptotes are independent of σ. For values of R somewhat greater than p the ξ values increase very slowly with σ which implies that large failure severities do not greatly degrade the performance of a relatively reliable system. In general, similar measures of failure severity should be investigated in order to establish the desired separation of the $A(\theta)$ and $F(\theta)$ distributions.

From numerical examination of the function it can be shown that the form of $F(\theta)$ is most important in the neighborhood of $\theta < 2\,\theta_{pA}$. It also can be shown that if $R > p$; then the exact form of $F(\theta)$ outside of $1 \leq \xi < \xi^*$ is of no consequence to system performance. Also the part of the distribution function of most interest lies in the range $0 \leq F(\theta) < p$ and this range can be bounded from above and below by the circular normal distributions of the type considered above. Insofar as subsystems are independent, the determining of $A(\theta)$ and $F(\theta)$ distributions as measured is applicable to subsystem testing. Of course, at subsystem testing one is faced with the problem of ascertaining the appropriate measures of performance which retain their appropriate meanings when the subsystem is integrated with the over-all system.

Unfortunately, obtaining the $A(\theta)$ and $F(\theta)$ distributions is not always a simple chore. Although failure data are collected regularly, it is not usually accompanied with a measure of the possible loss of accuracy or increase of miss distance or other system measure. In order to implement the computation of these distributions some evaluation of the performance must be obtained for each failure. This can be done by a subjective evaluation by an engineer, a computer simulation of the system operating with failures, or in remote cases, by instrumentation on the test vehicle. In some manner an estimate of the loss of performance must be added to each malfunction report.

The underlying theme throughout this approach is to partition reliability and systems performance into two separate categories in order that an evaluation can be more easily managed. Decisions about reliability and performance can then be accomplished in an orderly manner, one separate from another. For over-all performance measure, one can recombine reliability inherent accuracy, and failed accuracy. It may be reflected that a certain attitude is present in such an approach: let each factor assume its proper role and knowing the proper relationships among factors, good evaluation decisions can be made.

CHAPTER 4

A SUGGESTED APPROACH FOR PRACTICAL PERFORMANCE EVALUATION

Brief

The Pearson system of classifying frequency distribution curves obtained from empirical data is presented as a method for analyzing the $A(\theta)$ and $F(\theta)$ distributions. An example problem illustrating its application is presented.

Basic Comparisons to be Made

Although several interesting conclusions can be drawn from variations of the basic expression given in Equation (3-2), the practical problem of evaluating the $A(\theta)$ and $F(\theta)$ distributions from empirical data remains. The test data available for this evaluation will most likely be in tabular form and it will be necessary to summarize these data in order to measure the progress of the system's development. Before discussing these comparisons it is convenient to introduce subscripts to denote the distributions specified as design objectives and for different evaluation intervals. Let the $A_o(\theta)$ and $F_o(\theta)$ distributions represent the design objectives for the system. As previously noted the $A_o(\theta)$ distribution would be the design accuracy of the system and $F_o(\theta)$ the failed accuracy of the designed system which is not usually a specification. $A_i(\theta)$ and $F_i(\theta)$ would be the distributions obtained at the end of the i^{th} evaluation interval during the development testing period.

The five basic comparisons to be made are:

(1) $A_o(\theta)$ with $A_i(\theta)$: to determine how well the system has met its design objectives with respect to inherent accuracy

(2) $F_o(\theta)$ with $F_i(\theta)$: to determine how well the system has met its design objectives with respect to failed accuracy

(3) $A_{i+1}(\theta)$ with $A_i(\theta)$: to measure the relative shift in inherent accuracy performance between evaluation periods

(4) $F_{i+1}(\theta)$ with $F_i(\theta)$: to measure the relative shift in failed accuracy between evaluation periods

(5) $A_i(\theta)$ with $F_i(\theta)$: to measure the difference between the inherent and failed accuracy distributions.

Comparisons (1) and (2) can be approached as tests of hypothesis or "goodness of fit" tests. That is, for a given significance level, the null hypotheses are that $A_i(\theta)$ equals $A_o(\theta)$ or $F_o(\theta)$ equals $F_o(\theta)$.

The Pearson System

Unless nonparametric techniques are used it is necessary in Comparisons (3), (4), and (5) to first determine whether each of the distributions being compared is of the same type. As a start in resolving this problem it is suggested that Karl Pearson's[4] system of equations be used for classifying the distributions and for fitting distributions to the data. It is believed that, since there are no unified physical laws underlying the development process, any reasonable method for describing and scaling the process would suffice. Pearson's system of classifying frequency distributions is based on certain families of solutions to a particular differential equation:

$$dy/dx = (x + a)y/bx^2 + cx + d . \qquad (4\text{-}1)$$

The equation was obtained by putting dy/dx equal to the slope of a straight line joining two successive points of the discrete hypergeometric distribution. The Pearson system uses the coefficients of skewness and kurtosis in order to identify rapidly the general type of distribution. It should be noted that the coefficient of skewness is the ratio of the square of the third moment about the mean to the cube of the second moment or

$$\beta_1 = \mu_3^2 / \mu_2^3 \ . \qquad (4\text{-}2)$$

where μ_r is the r^{th} moment about the mean. Also the coefficient of kurtosis is the ratio of the fourth moment about the mean to the square of the second moment

$$\beta_2 = \mu_4 / \mu_2^2 \ . \qquad (4\text{-}3)$$

In using the Pearson system it is assumed that the first four moments as computed from raw data can be substituted for their respective population values. Pearson classified the solutions of the differential equation into 12 families of curves, those of one family being called Type I, those of a second Type II, etc.

On the basis of the coefficients of skewness and kurtosis, a valuable insight can sometimes be gained about the possible type of distribution which can be fitted to the empirical data even if Pearson curves are not used. Figure 8 is a plot of β_1 versus β_2 indicating the various areas and points which represent the main Pearson types of distributions and where other distributions are found on the plot. The following equations are for bounding curves on Figure 8 [see Reference (9)].

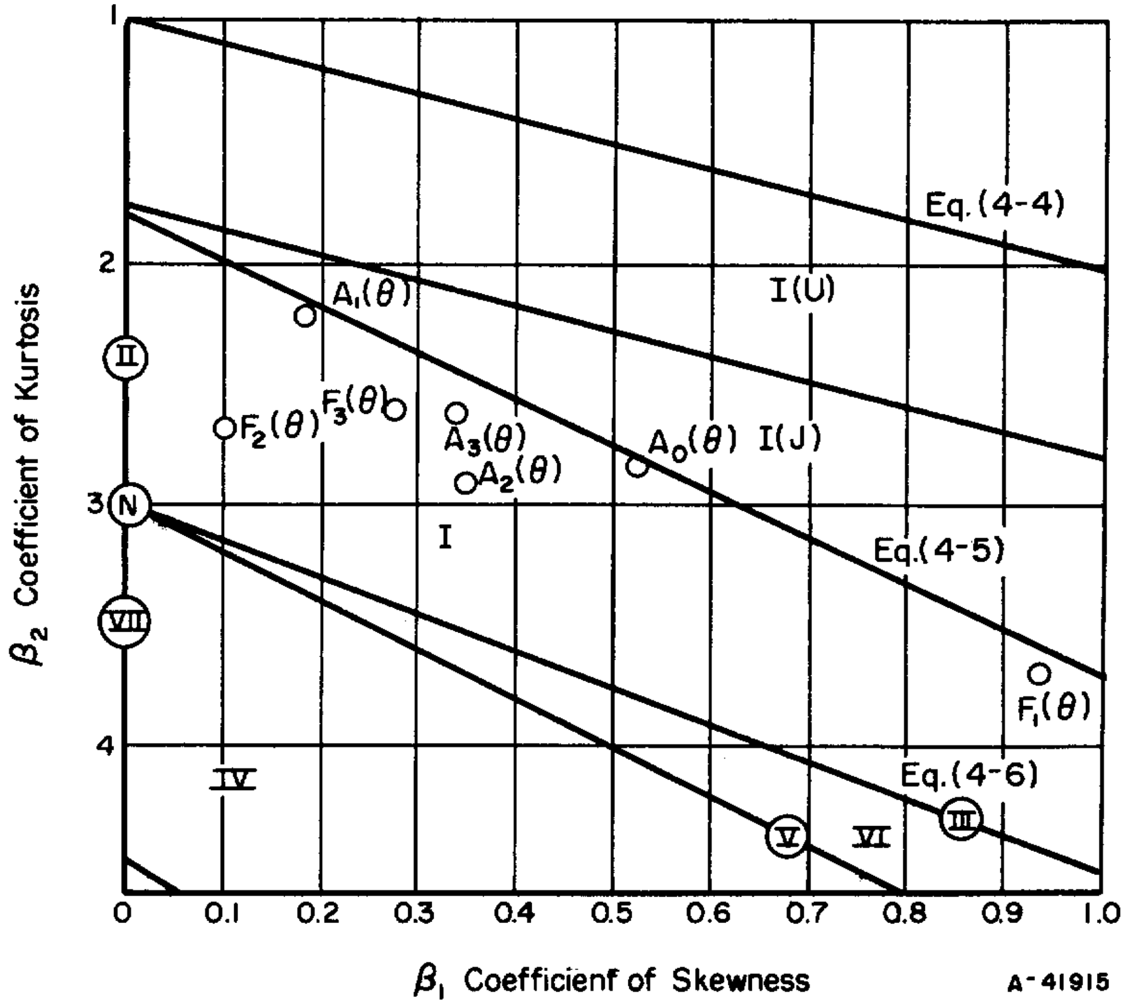

FIGURE 8. β_1 VERSUS β_2 FOR SAMPLES OF THE A(θ) AND F(θ) DISTRIBUTIONS

Upper limit for all frequency distributions:

$$\beta_2 - \beta_1 - 1 = 0. \tag{4-4}$$

Boundary of Pearson Type I (J-shaped area):

$$4(4\beta_2 - 3\beta_1)(5\beta_2 - 6\beta_1 - 9)^2 = \beta_1(\beta_2 + 3)^2(8\beta_2 - 9\beta_1 - 12). \tag{4-5}$$

Type III line:

$$2\beta_2 - 3\beta_1 - 6 = 0. \tag{4-6}$$

Type V line:

$$\beta_1(\beta_2 + 3)^2 = 4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6). \tag{4-7}$$

The line on and below which the third moment about the mean is infinite (for Pearson curves):

$$8\beta_2 - 15\beta_1 - 36 = 0 \tag{4-8}$$

The point (3, 0) in the figure represents the normal distribution. Additionally it may be noted that gamma distributions are essentially Type III and fall on a line defined by Equation (4-6). Beta distributions are essentially Type I, but are Type II when both parameters of the distribution are equal. The F distribution is a particular case of a Type VI, the t distribution is a particular case of a Type VII, and the "chi-squared" distribution is a particular case of the Type III. Thus the Pearson system and the knowledge of its relation to particular distributions can be useful in fitting sample data. The process of fitting a smooth curve to a sample does not add any more information about the sample distribution than that already contained in the sample. However, it can provide access to tables for simpler computations and standard statistical methods. Tables have been prepared for use with the Pearson system and are available in Reference (9).

An Example

To exemplify use of the Pearson system the following steps may be used as a guide. For both the $A(\theta)$ and $F(\theta)$ distributions the first four moments about the mean are computed and from these the coefficients of skewness and kurtosis are computed. After obtaining these coefficients a glance at Figure 8 will give an indication of possible distributions (e.g., normal) which could be used in place of the tables already prepared for use with the Pearson system. Reference (9) provides tables for $P = 0.005$, 0.01, 0.025, 0.05, 0.95, 0.975, 0.99, and 0.995 given β_1, β_2 for selected pairs of β_1, β_2 in the range $(-1 \le \beta \le 1;\ 0 \le \beta_2 \le 5)$. Of course with modern computing machinery other fractiles or alternative methods are sometimes as accessible as tables. However, for purposes of this example, assume that it is desired to use the tables.

In order to provide a numerical example, seven sets of random samples were drawn from a circular normal distribution, $1 - \exp\theta^2/\sigma^2$ for σ^2 equal to values shown in the first column of Table 2. Table 2 presents the results obtained from these synthetic data. If the development of the system were to progress in a manner similar to that shown in Figures 6 and 7, then it might be assumed that the $A_o(\theta)$ distribution corresponded to the distribution with a $\sigma^2 = 0.316$ and the distributions for other intervals as shown in Table 2. The coefficients of skewness and kurtosis have been indicated on Figure 8. On the basis of the tables in Reference (9) the 5 per cent fractiles are obtained in a straightforward manner. As is typical with statistical data the computed values do not correspond with true population values. However the general tendency for the increasing separation of the $A(\theta)$ and $F(\theta)$ distributions is reflected. It should be noted that there is no apparent tendency for the $A_i(\theta)$ fractiles to approach the $A_o(\theta)$ fractile. When faced with these types of data in an actual evaluation, further statistical tests or other types of evaluations would be in order.

TABLE 2. RESULTS OF THE EXAMPLE PROBLEM

	σ^2	β_1	β_2	$X_{.05}$	$\theta_{.05}$ Computed	$\theta_{.05}$ Theoretical
$A_o(\theta)$	0.316	0.5125	2.797	1.29	0.186	0.124
$A_3(\theta)$	0.448	0.3188	2.661	1.409	0.177	0.148
$A_2(\theta)$	0.632	0.3517	2.924	1.403	0.205	0.176
$A_1(\theta)$	0.775	0.1857	2.197	1.362	0.171	0.194
$F_1(\theta)$	0.894	0.9397	3.628	1.194	0.193	0.209
$F_2(\theta)$	1.000	0.1006	2.664	1.1532	0.316	0.221
$F_3(\theta)$	1.414	0.2710	2.599	1.400	0.332	0.263

Also in a practical problem the number of samples comprising each $A(\theta)$ or $F(\theta)$ distribution is going to depend upon the number of successful and unsuccessful tests. The uncertainty associated with any statistical test is going to be inversely proportional to this number. This can be seen immediately if one attempts to establish confidence intervals for statistics of either distribution.

The Pearson system is only one of many techniques available for scaling the evaluation of development progress. The individual circumstances in each evaluation case may dictate much about the statistical techniques used.

CHAPTER 5

CONCLUSIONS

This research has investigated the relationship between malfunction data, particularly human-initiated malfunctions, reported during development testing and system performance evaluation and reliability prediction. In addition the research effort was concerned with determining during development testing whether or not the system under test conforms to its design objectives.

During the research several basic points evolved which must be satisfied before system evaluation can be considered:

(1) Standards must be established for covering reporting methods for human-initiated malfunctions. Biases due to different individuals and concepts of human-initiated malfunctions should be reduced.

(2) Classifications and definitions must be established in order that human-initiated malfunctions can be appropriately identified and categorized. The equivalent of uniform inspection procedures and uniform sensitivity of reporting is necessary before meaningful data can be collected during development testing procedures.

(3) The interaction of the human factor activities and other development activities should be examined and recognized.

(4) Human factors specifications should be established.

(5) Management support must be given to human factors groups working directly on weapon system development.

Subsequent to the accomplishing of the above, analysis of system performance and reliability can proceed. It is set forth in the report that many human-initiated malfunctions occur early in development and are comparable to equipment parts with short mean lives. In a properly developing system these are gradually reduced until a random human error component remains. Human-initiated malfunctions must be interpreted in the light of over-all system reliability because they are part of failure data used to compute estimates of reliability. During development, reliability growth is often overshadowed by an apparent increase in failures due to the increasing number of tests and systems under test. A true determination of reliability growth during system development can be helpful in judging the role of human-initiated malfunctions.

Because of reliability growth during development, it is recommended that malfunction data be partitioned into reliability and performance categories in order to determine whether or not a system conforms to a design specification. In so doing it is possible to examine the inherent performance of a system with the failed performance or that system performance which remains upon failure of a component or subsystem. By finding the inherent and failed performance distributions and monitoring their changes during development, the true effect of failures, including human-initiated malfunctions, can be assessed. As development progresses the inherent performance, as determined from successful development tests [the $A(\theta)$ distribution] should approach the performance distribution as specified by design objectives. A measure of the separation between failed and unfailed performance distributions is also suggested as a measure of system development. For some distributions this separation can be measured by failure-severity parameters. Over-all actual performance may be measured by appropriately combining reliability and the failed and inherent performance distributions. Steps outlining the entire procedure are given in Table 3. Application of the methodology set forth would depict the role of human-initiated malfunctions in development and also provide an indication of whether a system was appropriately approaching its design objectives during development.

TABLE 3. STEPS TO FOLLOW TO MONITOR SYSTEM PERFORMANCE

(1)	Establish necessary standards for reporting malfunction
(2)	Collect data for each test
(3)	Associate value measure (e.g., accuracy) with each failure and for each successful test
(4)	Compute reliability in terms of the probability of no component failure
(5)	Fit growth curve to reliability or hazard rate data
(6)	Fit $A(\theta)$ and $F(\theta)$ distributions using standard statistical techniques
(7)	Compute failure severity parameters, if possible
(8)	Solve for performance function [Equation (5-2)]
(9)	Compare reliability with previously computed reliability, $A(\theta)$ with design performance, $F(\theta)$ with previously computed $F(\theta)$ and with currently computed $A(\theta)$
(10)	Evaluate results

REFERENCES

(1) Barlow, R. E., and Marshall, A. W., "Probability Distributions with Monotone Hazard Rate", Handout at Annual Meeting of the American Statistical Association, December 27-30, 1961.

(2) Boyle, R. S., and Nishime, F. S., "Sequential Analysis Methods Can be Used to Monitor Reliability Growth Curves", Proceedings of Symposium on Guided Missile Reliability, December 1959.

(3) Drose, J. S., "Failure Report Evaluation in Terms of Weapon Objectives", Proceedings of Symposium on Guided Missile Reliability, December 1958.

(4) Elderton, W. P., Frequency Curves and Correlation, Harren Press, Washington, 1953.

(5) Gumbel, E. J., Statistics of Extremes, Columbia University Press, New York, 1958.

(6) Kucij, T. C., "The Necessity and Philosophy of Failure Classification During Flight Tests of Developmental Guidance Systems", Proceedings of Symposium on Guided Missile Reliability, December 1958.

(7) Meister, D., "Personnel Logistics as a Factor in Achieving Weapon System Reliability", Proceedings, 6th Joint Military - Industry Guided Missile Reliability Symposium, pp 160-172, February 1960.

(8) Meister, D., "The Role of Human Engineering in the WS 107A-1 (Atlas) Reliability Program", Proceedings of Symposium on Guided Missile Reliability Part I, pp 245-254, December 1958.

(9) Pearson, E. S., and Hartley, H. O., Biometrika Tables for Statisticians, Vol I, University Press, Cambridge, 1958.

(10) Roberson, R. E., "An Approach to System Performance Prediction", Journal of the Franklin Institute, Vol 268, No. 2, pp 85-105, August 1959.

(11) Shapero, A., and Bates, C., Jr., A Method for Performing Human Engineering Analysis of Missile Systems, WADD TR-59-784.

(12) Shapero, A., Cooper, J. I., Rappaport, M., Schaeffer, K. H., and Bates, Chas. Jr., Human Engineering Testing and Malfunction Data Collection in Weapon System Test Programs, WADD TR 60-36, February 1960.

(13) Story, A. W., Man-Machine System Performance Criteria, Operations Applications Office, Deputy for Technology, Electronic Systems, AFSC, USAF, Bedford, Massachusetts.

(14) Tufts Handbook of Human Engineering Data, Tufts College, Institute of Applied Experiment Psychology.

(15) Van Cott, H. P., Checklist of Human Engineering Evaluation Factors, "CHEEF 2", American Institute for Research, September 1956.

(16) Van Cott, H. P., and Altman, J. W., Procedures for Including Human Engineering Factors in the Development of Weapon Systems, WADC TR 56-488, October 1956.

(17) Weiss, H. K., "Estimation of Reliability Growth in a Complex System with a Poisson-Type Failure", Journal of Operations Research, Vol 4, No. 5, October 1956.

APPENDIX I

GLOSSARY OF SYMBOLS

a, b	=	arbitrary bounds of an interval
$A_o(\theta)$	=	inherent accuracy distribution as designated by performance specifications
$A(\theta)$	=	inherent accuracy distribution
$F(x)$	=	cumulative distribution function
$F(\theta)$	=	cumulative failed accuracy distribution
$f(x)$	=	probability density function
r	=	arbitrary parameter
R	=	reliability
$P\{X\}$	=	probability of X
t	=	elapsed time
z	=	hazard rate
β_1 and β_2	=	coefficients of skewness and kurtosis, respectively
α	=	parameter of circular normal distribution
$\delta(t)$	=	damage or wear function
θ	=	system error (e.g., miss distance)
θ_{pA}	=	the p^{th} point of the $A(\theta)$ distribution
θ_{pF}	=	the p^{th} point of the $F(\theta)$ distribution
μ_r	=	the r^{th} moment about the mean
ξ	=	θ_p / θ_{pA}
ϕ	=	parameter of circular normal distribution
σ	=	failure severity parameter

APPENDIX II

GLOSSARY OF TERMS

assignable cause – a factor contributing to the variation in quality that is feasible to identify

coefficient of kurtosis – a dimensionless measure of the peakedness of a frequency distribution obtained by dividing the fourth moment about the mean by the square of the second moment

coefficient of skewness – a dimensionless measure of skewness of a frequency distribution obtained by dividing the square of the third moment about the mean by the cube of the second moment

cumulative distribution function – if x is a random variable, then the cumulative distribution function of x is defined to be the function F such that for every real number t, F(t) is the probability that a given outcome of x will not exceed t

development – the specific project design work leading to manufacture, including necessary tests and research leading to efficient design methods

failed system accuracy – the probability that if one or more components of a system fail, then the performance of the system will be less than a predetermined value

failure data – field, laboratory, and life-testing reports citing equipment failures or human-initiated malfunctions

functional test – a procedure to determine whether or not an ordered sequence of tasks or operations conforms to a specified performance

growth curve – certain solutions to differential equations used to characterize the phenomena of growth

hazard rate – the instantaneous failure rate of an object at a certain instant, conditional upon the nonfailure of the object prior to that instant

human-initiated malfunctions – an equipment failure or unscheduled hold (delay) wherein the human component can be clearly identified as the causative agent in the immediate train of events leading to the equipment failure e.g., misassembly, mishandling, or misadjustment by the human operator inherent system accuracy – the probability that if no component of the system fails, then the performance (error) of the system will be less than a predetermined value

job description – identifies the scope of activity of a worker, lists the tasks comprising the job, and the worker specifications (e. g., experience) required for the job

kurtosis – the relative steepness of ascent in the vicinity of the mode in a frequency distribution; peakedness as opposed to flatness

life length – the length of time an object will survive (not fail) with a specified probability

Pearson system – a system of classifying frequency distributions based on certain families of solutions to a particular differential equation

probability density function – if x is a random variable, then the probability density function of x is defined to be the derivative of the cumulative distribution function F of x; at points where $F'(t)$ does not exist, $f(t)$ is not defined; an equivalent definition is

$$f(t) = \lim_{h \to 0} 1/h \, P\{t - h/2 < x < t + h/2\}$$

probability distribution – if x is the random variable denoting the outcome of a random experiment, then the probability distribution of x is a set function which assigns, for every a, b, to the interval between a and b the probability that a given trial in the experiment will yield a value of x between a and b

process specification – instructions for receiving inspection, assembly, and installation of equipment; a process specification usually contains information too detailed for inclusion on an engineering drawing

reliability – the probability that a given system will give satisfactory performance for a given period of time when used in the manner and for the purpose intended

skewness – nonsymmetry in a frequency distribution; the graph of the frequency distribution has a long tail extending to the right or left

statistical test – a procedure to determine whether or not observed values or quantities fit a hypothesis sufficiently well so that the hypothesis can be accepted

system – an assemblage of objects united by some form of regular interaction or interdependence; a system is often referred to as part of a universe and the system itself is composed of elements, operations or subsystems.

variance – the mean of the squares of deviations from the arithmetic mean; analogous to moment of inertia.