A Critical Discussion on Bath-tub Curve

Tan Cheng
Institute served: China Association for Technical Supervision Information (CATSI)
E-mail: honesty108@hotmail.com

Abstract

The typical bathtub curve and its "standard" shape have been widely accepted as an engineering tool in reliability management and training. However, some recent researches reveal that the typical theory of 'bathtub curve' is not so convincing in some situations. Therefore, a critical discussion is developed to research the bath-tub curve along three significant periods of bath-tub; i.e. infant mortality, useful life and wear-out. The discussion is not only focused on the performance at different period but trying to reveal some in-depth theoretical considerations and practical contradictions behind every specific distorted "bath-tub". At last it concludes that the true meaning of bath-tub curve is on the business practice of the real world rather than any simplified models.

Keywords: bath-tub curve, infant mortality, useful life, wear-out.

1. Introduction

'Failure, for most parts of an operation, is a function of time' (Slack, 2001). In many cases, plotting the failure rate against a continuous time scale, the results will constitute the so-called 'bath-tub' curve (see **Figure 1**). From its shape, the curve can be divided into three distinct zones or periods quite readily. These zones differ from each other in failure rate and in causation pattern as follows: infant mortality, useful life, and wear-out. However, some recent researches reveal that the typical theory of 'bathtub curve' is not so convincing in some situations. The purpose of this paper is to identify and critically discuss the practical methods to quantify and eliminate failure in each part of 'bathtub curve' by appreciating the way the failure rate behave in time. Further, to show how to combine these methods in a coherent quality & reliability programme plan to reduce the initial-failure rate, extend the useful life, limit the random failure rate and take necessary action before wear-out period.

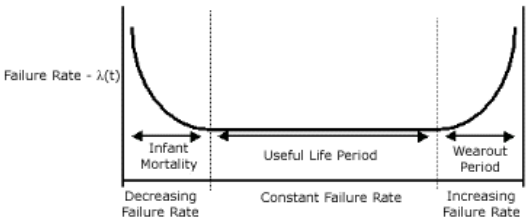


Figure 1

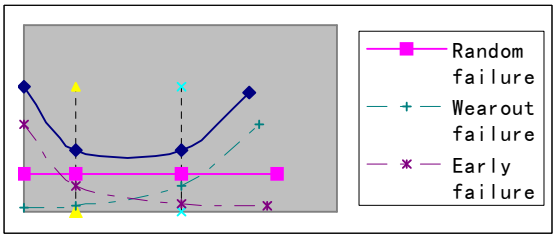


Figure 2

2. The Typical Theory of Bath-tub Curve

The typical theory of 'bathtub curve' has been widely accepted as an engineering tool. The bathtub shape is 'characteristic of the failure rate curve of many well designed products and components including the human body' (*Oakland, 1992*). The classic bathtub curve against time has three different periods: Decreasing failure rate for infant mortality; Constant failure rate for useful life; and increasing failure rate (without bound) for wear-out. Ebeling (*1997*) expresses this notion of bathtub curve as a composite of several failure distributions, and formulises it as 'a function of piecewise linear and constant failure rates'.

Cater (*1986*) however, argues this typical mode is only suitable for complex maintained components, which shows an additional constant failure rate for truly random failure, while for simple non-maintained components there is no such an additional constant failure rate. So there is a necessity to have a critical review on the three periods of the bathtub curve.

3. The infant mortality period

3.1 Identify and quantify failures in infant mortality

From the **Figure 2**, it can be seen that the initial 'infant mortality' period of bathtub curve is characterized by high failure rate and the items become less likely to fail as their survival time increases. For long time product, as what Juran (*1988*) has pointed that the 'infant mortality' is a major reason determining complaints rate. On the other hand, although the initial failure period is generally short, there are some exceptions. 'This period normally ranges from a few minutes to several hundreds hours'. So to provide customer good product reliability, it is necessary to reduce initial failure rate.

Generally different people have slight differences in the expression of the causations of initial failures. Ebeling (*1997*) suggests that poor quality control is one of major causations of early failures. The poor material is recognized by the 'Reliability Guidebook of Japanese Standards Association' as a major reason of failures in the infant mortality period. From Juran's (*1988*) view most of the failures during infant mortality period are results of identifiable causes, such as blunders in design, manufacture, or usage or of misapplication. This reflects Juran's quality philosophy, which is based on the presumption that the use of product can be fully understood. That's an important contradiction with Deming's quality thinking, who believes that the information sometimes is not only unknown, but inherently unknowable to us; and sometimes failures may be caused by 'Acts of God'. Interestingly, even Juran (*1988*) himself also admits that although reliability is determined by the quality of design, the achieved reliability is usually less than the attainable or intrinsic reliability because of 'the unanticipated environment during use, laps in quality of conformance, inadequacies in maintenance, etc'.

Based on Juran's (*1988*) proposal, the quantification process of reliability involves four phases: establishment of objective, apportionment, prediction and analysis.

—— **Establishing the requirements** here implies determine an ultimate acceptable failure rate before the product population leave the factory. This phase requires designers to understand the design in greater depth to define environment conditions and successful product performance precisely, especially process wording reliability requirements concerns to product specification. Moreover, Juran (*1988*) reminds the further critical wording of what constitutes a failure and the exceptions to the failure definition, which 'can readily be decisive as to whether the product meets the test'. Obviously, the failure definition is critical to the provisions of warranty for customers.

—— **Apportionment or budgeting** implies allocating the overall numerical reliability objective for each portion of the design, which makes up the total product

collectively. Additional, in this process engineering judgement with knowledge and previous experiences should be combined.

—— 'Inherent in the establishment of reliability requirements is the need to **estimate or predict reliability** in advance of manufacturing the product' (*Juran, 1988*). The prediction should be based on design information and past failure rate experience. In the final stage, as Juran (*1988*) has pointed, the prediction becomes 'the measurement based on data from field use of the product'. Through the prediction, a quantitative evaluation of the early failure rate can be achieved; the potential area for reliability improvement can also be identified. In addition, this phase can be effective by using some software. However, the people who process prediction must state the assumption behind their predictions, and further to provide 'a good quantitative measure of the uncertainties in the number' (*Editorials of IEEE transaction on reliability, 1999*).

—— Analysis phase here means to identify the strong and weak points for improvement, trade-off and some other actions. Some analysis techniques, such as FTA, FMECA, and 'worst-case analysis' can be adopted during this phase.

3.2 Eliminate failures in infant mortality

Ebeling (*1997*) suggests four methods to reduce failures during infant mortality period, they are: Burn-in testing or debugging testing, Environment Stress Screening, Quality control and Acceptance testing. In addition, Juran (*1988*) adds in accelerated testing as a way to identify and eliminate initial failures.

3.2.1 Burn-in or debugging testing

Due to the high failure rate in the initial failure period, burn-in (for electronic items) or debugging are widely accepted as an approach to screening out failures before they leave the factory until the product population reaches a low failure-rate. The failed products are scrapped (if not repairable) or minimally repaired (if repairable).

The application of debug or burn-in is to keep the weak unit failure takes place in the test period rather than in service. However, Kuo (*1984*) argues that the traditional decision criteria used to burn-in components is simple and arbitrary. These kinds of models cannot sufficiently reflect the real world. The editorial of IEEE (*1999*) comments that these simpleminded conceptual and ideal models 'bring mind some desired utopian state, not the rough-and-ready quick-and-dirty simple-mindedness of our equation'. It is ivory tower, not engineer's tool. Further, Juran (*1988*) also warns some risks of testing, such as intended use differs from actual use; model construction is not flexible enough for subsequent production, the small number of models building is not adequate for mass production; and subjective evaluation of test results because of pressures to release a design for products. So the deeper studies of the actual

conditions of use and engineering knowledge should be involved into design.

Another major problem associated with burn-in is to decide exactly how long and at what level of assembly should the components be burned in, trading off appropriately the need of reliability and the total cost. The insufficient burn-in may result in high initial failure rates, which may increase high field repair costs. On the other hand, with excessive burn-in, the decreased failure rate may cause increased cost and recurring costs. However, in some situations, burn-in or debugging may be not needed at all, such as for some inexpensive components, where 'repair is simple and the consequences of failures are trivial' (*Kuo, 1984*). But the debugging or burn-in will be paid eventually, directly or indirectly by the customers. The customer will trade off the costs of burn-in against the costs of receiving unreliable products. So there are three ways to balance this problem: (1) to minimize the cost of customer and (2) to maximize the benefits of testing (3) to trade off the cost and benefits of burn-in testing.

Plesser and Field (*1977*) considered that the optimising debugging or burn-in time could minimize the cost of ownership. Cheng' studies also support this opinion. He determines the optimal burn-in time by minimizing the system cost, for example to reduce costly early field replacements.

Cozzolino (*1966*), on the other hand, adopted a test policy 'for determining the sequence of decisions that would maximize the excepted benefits of testing for initial defects' (*Ascher, 1984*). Chandrasekaran has similar consideration. He recognizes the problems of minimizing life-cycle cost and determines the optimal burn-in time so as to maximize 'mean residual life' (MRL) and to realize the total cost optimisations. With respect to this point, Park (*1985*) reminds that although it is popular believed that with the increase of burn-in period, the failure rate of product surviving the burn-in tends to decrease until reach the useful life period where failure rate is constant, yet this popular belief is questionable. He chooses the 'mean residual life' (MRL) as a parameter to study the effects of burn-in and concludes that in the post-burn-in period, the 'time at which a bathtub failure rate is minimum does not maximize the MRL (*Park, 1985*), and the MRL in the constant failure-rate region of a bathtub curve is not constant too.

Marko and Shoemaker's research of optimal burn-in solution is based on the definition of life-cycle cost. They regard the life-cycle cost as a combination of burn-in cost and field failure cost. Different from the two methods mentioned above, they use 'exhaustive search techniques to optimise space module burn-in and thereby minimize a part of the life-cycle cost' (*Kuo, 1984*). The optimal solution is then identified by comparing the increased costs associated with added burn-in against field saving from failure reductions.

However, this notion of trade-off between quality and cost is not accepted by all the

quality gurus. Juran believes COQ can also be used to establish the goal of quality programs: 'to keep improving quality until there are no longer positive economic return'. This analysis is built on the assumption that as defects become fewer and fewer or reduced to lower or lower level, the cost of prevention and appraisal cost will approach to infinity. Then, the minimized cost of quality can be achieved at the point where additional spending on prevention and appraisal equals to the savings resulted from these preventive and appraise activities in failure cost. On the contrary, Deming believes that that lower cost can be achieved through improved quality. As what Slack (*2001*) has pointed Deming's basic philosophy is that quality and productivity can increase simultaneously by decreasing the 'process variability'. He called in question that the traditional view of the existed trade-off between quality and productivity. He thinks costs and quality are not to be trade off against each other. The reduced rework, fewer delays, better utilization of resources, lower cost, happier people on the job, more jobs and hence improved market share and long term business survival can be achieved from the improved quality.

3.2.2 Environment Stress Screening

Both Juran (*1988*) and Ebeling (*1997*) think 'Environment stress screening' should be used to eliminate the early failures due to weak parts, workmanship defects. Juran (*1988*) define it as tests performed at lower level of the product. Further, he also suggests combine it with life testing to identify weak points in new designs. Moreover, Yan (*1997*) also suggests the implementation of 'Environmental stress screening' in eliminating patent failures during useful life period, which will be discussed later in **Clause 4.4**.

3.2.3 Quality control

Quality control is concerned with identifying and controlling product and service characteristics and further to prevent the occurrence of failures. For example, process control chart can be used to detect potential problems when a process was going out of control and then take corresponding actions before the occurrence of failure.

3.2.4 Product reliability acceptance test

From Juran's (*1988*) point, acceptance tests can be used as 'periodic evaluations of reliability of production hardware', especially when design, tooling, processes, parts, or other characteristics have some changes.

3.2.5 Accelerated testing

Accelerated testing is 'a common form of securing reliability test data at reduced testing cost' (Juran, 1988). In accelerate testing products will burden abnormally high

level of stress and / or environment to make them fail sooner. The short life under severe conditions then can be transferred into normal conditions by extrapolation.

However, the shortcomings of accelerated testing are obvious. Firstly, in practice the time of accelerated testing is hard to be controlled to ensure it is properly correlated to normal use time and avoiding overstating the expected life. Secondly, accelerated testing may introduce some new failure modes, which do not occur under normal conditions. Further, these new failure modes may result in costly redesign, 'which provide no benefit from the standpoint of the original product requirements' (*Juran, 1988*). Generally, although the misled risks of accelerated testing do exist, the benefits from testing can be substantial. So Juran (*1988*) suggested that the involvement of engineering judgement in such tests is crucial.

4. The constant failure rate period

4.1 General explanation of 'useful life' period

The infant mortality period is followed by a nearly constant failure rate period known as 'useful life'. In this region failures occur by purely chance. Additional, as the failure rate is constant, this is the only region, in which the exponential distribution can be valid and thereby the time between failures is exponentially distributed.

4.2 Identify and quantify failures in useful-life

Juran (*1988*) thinks that the failure of this period result from the inherent limitation of design and accidents caused by usage or poor maintenance. Additional, Ebelin (*1997*) adds the environment random load and 'Acts of God' into causations.

On the other hand, as the failure rate during this period is constant, the failure rate follows exponential distribution (or Weibull distribution with the shape parameter of $\beta = 1$). The mechanism of exponential distribution, as Juran (*1988*) has pointed, is that of random failures which are independent of accumulated life and consequently are individually unpredictable. This mechanism is important for the calculation system reliability, which is often assumed that the 'system reliability is the product of the individual reliability of the n parts within the system' (*Juran, 1988*). That is so-called 'product rule':

$$P_s = P_1 P_2 ... P_n \qquad (1)$$

As the 'product rule' also assumes (1) the failures of any parts will cause system failure, and (2) reliability of the parts are independent of each other, so when each parts follows exponential distribution, then:

$$P_s = e^{-t_1\lambda_1} e^{-t_2\lambda_2} ... e^{-t_n\lambda_n} \quad (2)$$

Further, if t is the same for each part:

$$P_s = e^{-t\sum\lambda} \quad (3)$$

Therefore, when the failure rate is constant, the system reliability can be predicted by addition of the part failure rate. Although the assumption is not always valid and the assumption of exponential distribution should be supported by data collection of failures, yet because of the simplicity of the exponential density function, it has been used extensively in reliability work.

But Jensen (*1990*) argues that in the bathtub pattern, the constant (or nearly constant) failure rate is found in strictly life-tests. During these tests the random fluctuation does not take place. Juran (*1988*) also admits that the assumption of constant failure rate is rightly questioned. But the 'experience suggests that this assumption is often a fair to make' (*Juran, 1988*). Further he believes that taking design actions to produce a constant failure rate is more fundamental than arguing the validation of assumption.

4.3 Eliminate failures in 'useful life'

From Juran's (*1988*) view, good control on operation and maintenance procedures can be applied to eliminate the accident failures during useful life period, which are caused by poor maintenance or usage. However, the basic reduction of failure rate requires a basic redesign. Differently, Ebeling (*1997*) suggests using redundancy to reduce failures during 'useful life'. Ascher (*1984*) argues that these kinds of contradictions may result from the ignorance and misunderstanding of the existence of 'two bathtub curve', one is 'a model for a population of parts, and another bathtub curve with a different interpretation is a model for a system'

In addition, it is worthy of attention that the decrease in failure rate during 'useful life' period (or increase in MTBF for repairable component) does not result in a proportional increase in reliability (the probability of survival). For example, in equation $R = e^{-\lambda t}$, if t = 1h. A fivefold increase in MTBF from 20 to 100h can only produce 4% increase in reliability. But by increasing MTBF from 5 to 10h, 8% increase can be gotten in reliability. The understanding of this exponential character is important, because the failure rate or MTBF are always used as criteria for decision-making, which will affect reliability, whereas 'the probability of survival for a specified time t may be the more important index to the customer' (Juran, 1988)

4.4 Some specific bathtub curve

Different types of component can exhibit significant variations on this basic bathtub

shape. Some of the differences represent in useful life period. Billington's (1992) believes that two particular examples cover the two extreme cases, they are: electronic component and mechanical component. Comparing to mechanical components, electronic components (See Figure 2) are usually associated with a relative long useful life, while for mechanical components (See **Figure 3**), this period is very brief. Moreover, Slack (*2001*) also points that for the failures of operation which rely more on human resources than on technology, the useful life of bathtub curve is not evident and the initial failure period may be followed by a long period of increasing failure. (See **Figure 4**)

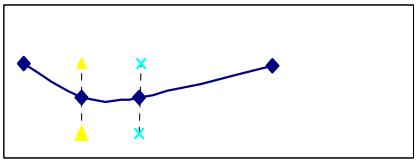


Figure 3

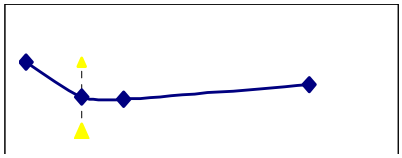
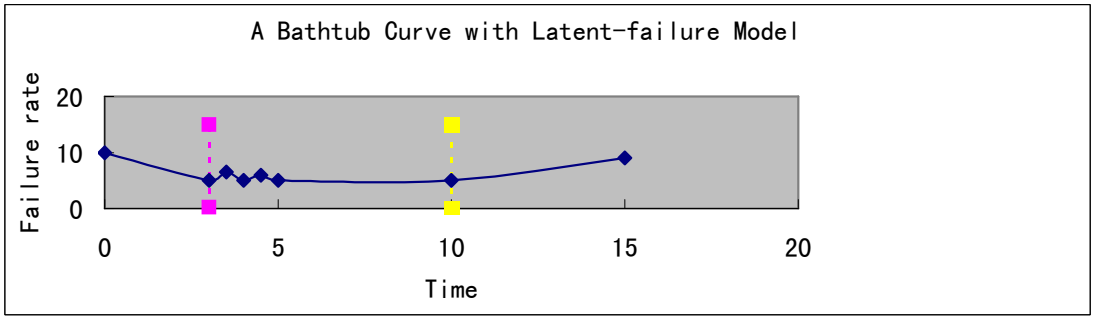


Figure 4

Moreover, some recent research in the electronic industry has identified another failure pattern: latent failure. The causations of latent failure is the undetected failures, which 'can not be readily defected by inspection or functional testing until it is transferred into a patent defect by environmental stress applied over time' (*Yan, 1997*). The characteristic of latent failure is when the stresses, both from internal and external, exceed the strength of system; there will be a jump in the failure-rate curve because of the occurrence of latent failure. (See **Figure 5**)



FFigure 5

Further, Yan (*1997*) also argues that in conventional bathtub curve, the failure time is assumed to be continuous between zero and infinity. However, in practice, the useful life of most electronic system is finite. The technology obsolescence factor should be considered. Therefore, the combination of technical obsolescence and latent defects can be 'modelled as a truncated mixture distribution' (*Yan, 1997*). (See **Figure 6**) However, as Yan (*1997*) has pointed, although the actual shape of bathtub curve can be various, the general shape remains the same'.

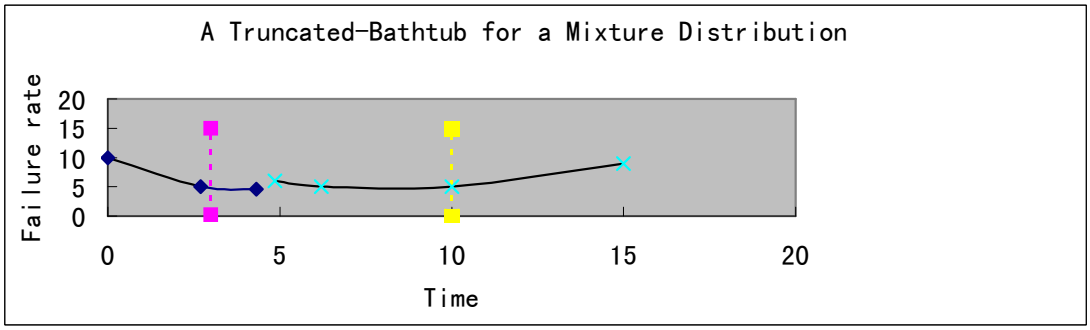


Figure 6

5. The wear-out period

5.1 General explanation of wear-out period

The wear-out period is characterized by a rapid increasing failure rate with time. From Billington's (*1992*) point of view, this period is more evident than the other two periods as in this region, the failure density function will increase firstly, and then decrease to zero for the obsolescence of components.

5.2 Identify and quantify failures wear-out period

In the wear-out period, the failure density function can often be represented or approximated by normal distribution. The gamma and Weibull distribution are also can be used to quantify failures during wear-out period. The reason is these distributions have shaping parameters, 'the variation of which can create significantly different characteristic shape' (*Billington, 1992*). Further, Billington's (*1992*) points that if the wear-out region is normal distributed 'the time at which it enters the wear-out region depends upon the standard deviation of the distribution'.

However, Billington (*1992*) also reminds that the reliability prediction based on 'useful life' failure rate is invalid and extremely optimistic to be used with wear-out period. On the other hand, although the 'useful life' failure rate is analytical simple and the 'useful life period can be extended by careful and regular preventive maintenance and replacement', In practice, the component may fail long before the MTBF or MTTF is reached owing to the wear-out.

5.3 Eliminate failures in wear-out period

Juran (*1988*) suggests that the failures during wear-out period are due to old age. From his view, to reduce the failures it is necessary to bring preventive replacement of the dying components into effects. Ebeling (*1997*) agrees with that and adds preventive maintenance as another method to reduce failures during wear-out period. Cater

(*1986*) further comments that for mechanical system with ideal maintenance, (that is all failed components will be replaced by identical components without disturbing the system) the failure rate of system can remain the steady state and then the wear-out section of the bathtub curve could be eliminated. But Cater (*1986*) also admits that in practice this 'ideal maintenance' does not practicable. On the contrary, effective wear-out can prevent the unnecessary outmoded or uneconomical repair of mechanical equipment.

5.4 Arguments about the wear-out periods

Although the wear-out period of the bathtub curve is widely accepted, some researchers yet call into question of the shape and existence of the wear-out section of the bathtub curve.

George (*2000*) argues that the nonparametric estimates of failure rate function can reveal some unexpected thing, which is 'the bathtub curve doesn't always hold water' (**Figure 7**). He reminds the attention to retirement and points that retirement, which here means 'fewer operating hours per calendar time unit' (*George, 2000*) can also causes the decrease of failure rate. The 'bathtub' doesn't hold water when the retirement occurred earlier than wear-out.

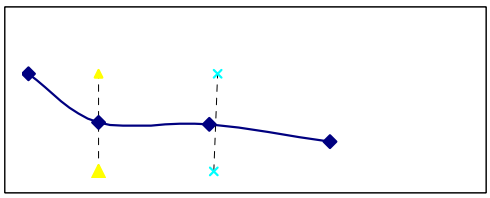


Figure 7

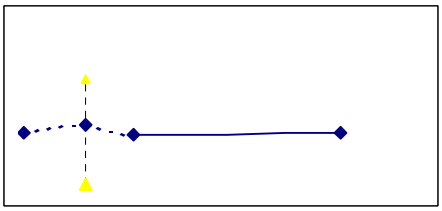


Figure 8

Similarly, Ascher (*1984*) also argues that although the bathtub curve theory represents the view of many reliability engineers and it is a better mode for some system, the repairable system bathtub curve contradicts the limitation theorems. The limitation theorem, which is illustrated in the **Figure 8,** indicates that the system will settle down to a constant value in the long life region, just where the bathtub curve forecasts that the failure rate will increase sharply. Further, as Ascher (*1984*) has pointed the reconciliation between bathtub curve and limit theorems will show a curve would show an inconspicuous wear-out period. In addition, Tsang (*1995*) declares that an extensive study in the airline industry proves that the bathtub curve is not a universal mode that applies to all items, and 'as much as 89 percent of all the airline equipment items do not have a noticeable wear-out region throughout their service life' (*Tsang, 1995*).

6. Conclusion

The bathtub curve is a common model to describe the failure rate of a population and

has been widely accepted as an engineer tool. However, as discussed above in some situations the typical bathtub theory is not so convincible. As what Warrington (*2001*) has suggested that the simplicity of definition is its downfall. So it's really necessary to combine the methods, which is used to identify and eliminate failures together and implement them as a coherent quality and reliability programme aiming at actual required function, stated conditions and specific period, associated with consideration of costs, requirements of customer, design, business and manufacturing practice as well.

Acknowledgement

ALL FOR TINA

Bibliography

1. Ascher, H (1984) 'Repairable System Reliability – Modeling, Inference, Misconceptions and Their Causes', 1st, Marcel Dekker, Inc. New York.
2. Billinton, R (1992) 'Reliability Evaluation of Engineering System', 2nd, Plenum Press, New York and London
3. Carter, A.D.S. (1986) 'Mechanical Reliability'2nd, Macmillan Education LTD, London
4. Ebeling, C.E. (1997) 'An Introduction to Reliability and Maintainability Engineering', International ed. Mcgraw-hill, Inc, New York.
5. Editorial 'Models for Quality & Probability', IEEE Transactions on Reliability, VOL 48, No.2, 1999June
6. George, L (2000) 'The Bathtub Curve Doesn't always Hold Water', <http://www.asq-rd.org/articleBathtub.htm>, [accessed 18, February, 2001]
7. Jensen, F (1990) 'Burn-in: An Engineering Approach to the Design and Analysis of Burn-in Procedures', 7th, John Wiley & Sons, Chichester
8. Juran, J. M (*1988*) 'Juran's Quality Control Handbook', 4th, McGraw-Hill Inc, London
9. Kuo, W (1984) 'Reliability Enhancement Through Optimal Burn-In', IEEE Transactions on Reliability, VOL R-23, No.2, 1984 June.
10. Leitner, P. M (1999) 'Japan's post-war economic success: Deming, quality, and contextual realities', *Journal of Management History*, Vol. 5, No. 8, <http://www.ebsco.com/online/direct.asp?ArticleID=5775AVG6PQY2EYU19H5E> , [accessed 18, February, 2001]
11. March, A (1996) 'A Note on Quality: The View of Deming, Juran, and Crosby', *IEEE Engineering Management Review* Vol. 24, No.1.
12. Oakland, J. S (1992) 'Total Quality Management' 2nd, Butterworth-Heinemanm, Oxford
13. Park, K.S. (1985) 'Effect of Burn-In on Mean Residual Life', IEEE Transactions on Reliability, 1985 September
14. Slack, N (2001) 'Operations Management', 3rd, Prentice Hall, London
15. Straker, D (2001) 'What is quality' *Quality World*, Vol. 27, Issue 4.
16. Silvestro, R (1999) 'Deming', *Encyclopedic Dictionary of Operations Management*,

1st, Blackwell Publishers Ltd, Oxford.

17. The Japanese Standard Association 'Reliability Guidebook' 3rd ed. Nordica International Ltd, Hong Kong.
18. Tsang, A (1995) 'Condition-based maintenance: tools and decision making'. *Journal of Quality in Maintenance Engineering*, 8/1/95 (Vol. 1, No. 3), http://www.ebsco.com/online/direct.asp?ArticleID=852Y2CX3MY6GAJ24VRR9 [accessed 18, February, 2001]
19. Yan, L (2000) 'Economic Cost Modelling of Environmental-Stress Screening and Burn-In', IEEE Transactions on Reliability, VOL 46, No.2, 1997 June.
20. Warrington, L (2001) 'WMG Notes of Reliability Programme Management' Section 9, part 1