QUALITY AND RELIABILITY ENGINEERING INTERNATIONAL, VOL. 10, 391–398 (1994)

PRACTICAL RELIABILITY TESTING CONSIDERATIONS

SAMUEL J. KEENE AND PAUL VATTANO
Storage Technology Corporation, 2270 South 88th Street, Louisville, CO 80028-0001, U.S.A.

SUMMARY

This paper covers practical test concerns such as proper test design and analysis of data. It will speak to software and hardware testing as well as test design techniques such as design of experiments. Reliability and test terminology will be given. Some test lessons learned will be shared.

KEY WORDS Testing Test techniques Reliability growth Design of experiments Weibull analysis

INTRODUCTION

Why do we test our products? It is most often to verify that they meet the requirements and expectations of the customer. Testing typically exposes faults or bugs that have escaped earlier bug removal steps such as design reviews, walk-throughs and inspections. In this regard testing becomes a test and fix process where faults are discovered and then removed from the design. This process results in reliability growth of the product under test.

It is often said that: 'reliability cannot be tested into a product'. Reliability is a design attribute that is achieved during design, but it is also true that testing successively refines the product. This refinement enhances the product reliability as seen by the user.

The reliability growth of a fielded software product is shown in Figure 1. The heuristic that is often used is that the software has only 10 per cent of the faults remaining after 48 months of field experience. Testing produces a similar growth curve. Software subject to test acceleration (such that the bugs come down faster) may reach a comparable improvement in 6–9 months of testing.

Testing generally serves two basic purposes. First, it can be used to refine the design. It is finding faults and failure susceptibilities in the code. This is a test and fix situation (TAF) where the code is successively refined by detecting and removing faults. Testing is finding the stress points in the design. There is a fundamental rule in testing:

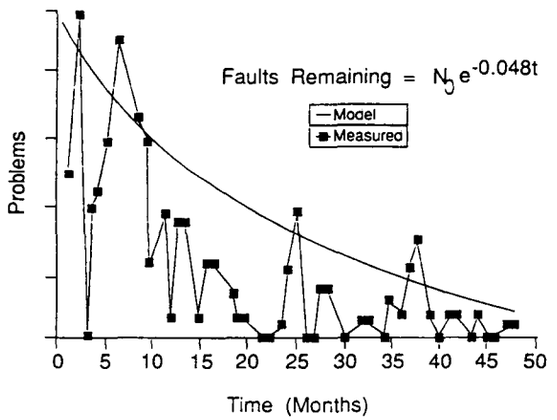

Figure 1. Reliability growth of fielded software

> Bugs in a program hate to be alone

So where one bug is found is a good place to look for another. There is abundant field evidence that fault distributions follow the Pareto principle. That is, the predominant number of software faults, for instance, are typically found in only a few modules. Software testing and reliability is much in the current focus. There are valuable concepts currently developing in this field. These concepts will be discussed in the fault/failure section below.

FINDING SOFTWARE FAULT(S)

In testing we are attempting to locate any faults in the design and remove them. Alternatively we are attempting to verify the absence of faults or ensure that they are sufficiently few that the product will meet its specified life and reliability requirements. Faults are the inherent failure suceptibilities in the design. Some of them will probably never be instigated in the fielded product. They are too illusive and will not be subjected to narrow triggering conditions needed to activate them and cause a failure. A failure is typically needed to signal the presence of a fault. This failure starts the sequence to identify, localize and remove the initiating fault. A failure is simply a departure from the customer's preference for how he desires the product to operate. A fault is a failure susceptibility. It can lead to failure given the proper operating stimulus.

Consider the simple Boolean addition operation:

$$A + B = C$$

This operation provides the intended results when A and B are 1 or 0. But how does this operation resolve negative numbers, irrational numbers and

CCC 0748-8017/94/050391-08

Received 10 January 1994

alphabetic characters? If the design is sufficiently robust, the operation will indicate an errant input condition. This would be the preferred resolution of this problem input. If these off-nominal inputs are not properly accounted for in the design, then there is a fault here, and failure can result. In the worse case, a bogus result would then be produced with no indication of the errant input.

Faults are logical concepts. The failure propensity of a fault may typically be removed in several ways. Recognize that there is not a unique physical place in the software or hardware that has to be changed to correct the fault. This flexibility of fault correction is illustrated frequently by system engineers correcting a hardware problem with a software fix. This is often done since software is easier and cheaper to change than hardware. Also, a single failure can lead to identifying multiple faults. This was the case in one of the classic space shuttle test failures. This software was tested more than 24 hours per day by running the code on multiple test stations. In one test situation, normal operations were temporarily suspended. When the astronauts returned, the system failed on restart. It was being restarted in an off-nominal manner that the software designers did not anticipate and the software could not handle. This was termed the 'multipass problem'. The fault condition was uncovered and the code structure identified that caused the failure. Then all 450 KSLOC of the flight code was parsed. A total of three instances of this fault (failure susceptibility) were located and corrected. Here one test failure led to the identification and removal of three faults in the code.

LIFE TESTING

The second main reason for testing, beyond purifying the design, is to measure the reliability of the product and its expected lifetime. This requires using some accelerated test means so the testing can be done in a useful time frame. Often, commercial products are designed to last 5 to 10 years. Testing usually runs less than a year and the goal is to certify that the product meets its specified life goals. So test acceleration is required. This system testing can estimate the life of the fielded product. Product refinement (TAF) as well as product life verification are both viable test concerns.

Typically, reliability specifications define a required percentage survial life at a specified confidence level. For example, one might want to ensure that 95 per cent of a population of mechanical switches will last for one million cycles. This is often called the L5 life since it is estimated 5 per cent or less of the population will fail in this time. One would have to test all possible samples (the universe of samples) to achieve 100 per cent confidence in such a prediction. Obviously this is impractical, so a trade-off is made to define the acceptable confidence level for this specification, of say 90 per cent. The specification would be made more stringent in the case of safety-critical applications where life was in jeopardy.

Test acceleration

There are a number of ways to accelerate failures and thereby shorten the time required for a reliability test.[1] If the component life is defined in relation to a number of cycles, the most obvious acceleration is simply to operate the equipment at a higher speed or higher duty-cycle than normal. In testing mechanical switches, for example, one could automate the test by using a high-speed actuator to speed switch actuations so that the test could be completed sooner. Sometimes higher speed operation increases dynamic loading or produces thermal effects, and these must be taken into account as accelerating factors in the analysis.

The other way to accelerate the test is to intensify an environmental stress or physical parameter that shortens life. Such stresses include temperature, vibration, humidity, voltage, current or loading. Common functions such as the Arrhenius relation or the power rule are used to relate the fractional life at stress to the nominal life using an acceleration factor (see Reference 1, pp. 71–107, and Reference 2). The life acceleration factor is usually defined in relation to one of the commonly used fractional lives such as L10. After the fractional life is determined at the accelerated condition (say by Weibull analysis), it may be calculated for the nominal application condition using a derived acceleration factor. Empirical acceleration factors can often be found in the literature, but sometimes the factors and the functional form of the acceleration must be determined in the life test itself. This requires multiple tests at different accelerating stresses. A larger sample size may be required for such tests. Data from such a test is depicted in Figure 2 (adapted from Reference 1, p.74).

Weibull data analysis

A useful life model for mechanical and electromechanical devices is the Weibull distribution (Reference 3, pp. 1–2). The Weibull distribution is

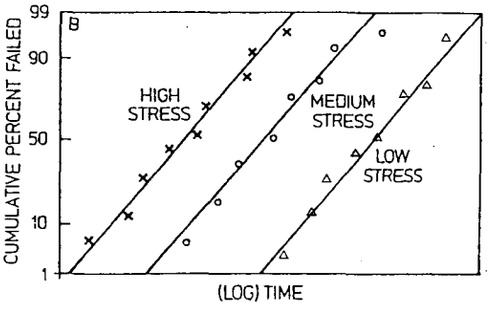

Figure 2. Typical accelerated test data

versatile and can represent the decreasing failure rates associated with early life or burn-in failures (slope <1). It can also represent the random failure modes where failure rates are found to be constant in time (slope =1). Lastly, it also can represent the wear-out case where the failure rate is increasing (slope >1). A large number of mechanical and electromechanical failure mechanisms can be fit to this distribution.

The Weibull distribution can represent distributional data as a straight line on Weibull probability paper (see Figure 3). The ordinate is the cumulative percentage of the sample failing, and the abscissa is time to failure. The slope, called the shape factor, is related to the nature of the failure mechanism (e.g. random or wearout). If the data is found to lie along a broken line (two intersecting lines) that would indicate that a change in failure mode had occurred. Random failures of some parts in early testing could be followed by wearout of other parts later in test.

In a typical Weibull test, a sample of parts is put under test and monitored for failures. Minimum sample sizes of 10 to 15 are required for significance. However, smaller sample sizes are often used when parts are unavailable, with a caution in the report that conclusions are largely qualitative. When failures occur, they are plotted on Weibull paper. Analysis of the plot can reveal multiple failure modes and pre-stressing of the samples (Reference 3, pp. 3–13). The fractional lives and confidence levels may then be determined (Reference 3, pp. 2–7). Software such as WeibullSMITH is available that will do the calculations and graphing.[4]

The required test time can be extensive to confirm L1, L5, or even L10 lives. Test time is inversely related to sample size, so as large a sample as possible should be used. Life testing should be started as soon as relatively stable prototype units are available to minimize delay in the product development process.

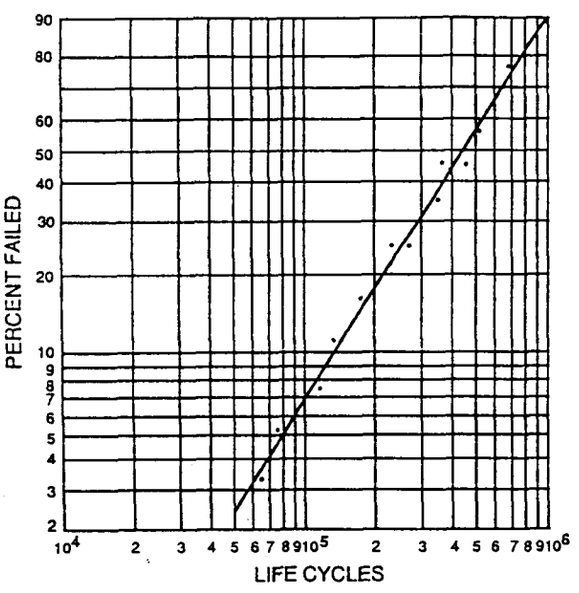

Figure 3. Weibull probability plot

Test accuracy and repeatability

In testing sample units, one typically characterizes samples prior to putting them on test. These measurements are repeated at intervals throughout the test to see if any changes have occurred. A necessary part of doing this is to first understand from the beginning of test the accuracy and repeatability of the data taken. Some basic testing safeguards are also suggested.

1. *Systematic error and instrumentation calibration.* For results to be meaningful, the errors introduced by the test instruments must be understood and reported. Test equipment errors are usually systematic errors, that is consistent positive or negative discrepancies from the true reading. If equipment is not calibrated to a traceable standard, these errors are unknown, and the test results are questionable. Traceable standards are those traceable to standards maintained by the National Institute of Standards and Technology (NIST), in the U.S.A., or those using a primary physical constant (e.g. the speed of sound in a particular substance). Most equipment manufacturers quote a systematic error and a maximum calibration time interval after which the systematic error is unknown. It is good practice to rigorously follow these calibration intervals and report systematic error for all tests.

2. *Random errors and repeatability.* Random errors are those errors in test observations that may be either high or low, unlike systematic errors which are consistently positive or negative. Examples of random error are small observation errors by the experimenter and small apparatus disturbances (as due to slight temperature fluctations). The random error is calculated after repeated measurements as a multiple of the standard deviation of the mean (depending on the confidence required). Obviously if the random error is large relative to the measured quantity, the test is not dependable and the causes of the errors must be studied and sufficiently reduced to an acceptable level.

The initial measurements on test units should be repeated several times to help to better size the measurement error. This can also identify any temporal drifts in the measurement apparatus. In one experience, one of the authors was measuring photocells prior to life testing. There seemed to be a shift in the measurement data over time. The data was repeated and the drift verified. These measurements were of all supposedly identical samples. The measurement apparatus was found to have a resistor value drifting during warm up that was causing a shift in the photocell readings. This resistor was upgraded with a power resistor and the measurements stabilized. Taking several sets of initial data at

different times and varied measurement sequences alerted us to this potential problem.

DESIGN OF EXPERIMENTS

The most effective and efficient test technique lies in the realm of design of experiments (DOE). Here one is looking at how the device under test will perform over its entire operating range. DOE takes a highly parallel, global look at the device under test (DUT). It alleviates the running and rerunning of many sequential tests. It also reveals application factor interactions which sequential testing cannot detect. DOE in its fullest application will involve representations of all the development and field product views. Thus it is a strong force to promote concurrent engineering.

Design of experiment—HE–NE laser

The greatest reliability improvement programme that one author has ever participated in or witnessed dealt with refining a Helium–Neon laser for a printer application. In the early 1970s we found the laser samples that we evaluated to be very erratic in their parameters, especially in terms of useful life. Our printer application demanded a life and power stability of the laser that was significantly beyond that which was available at that time. Our laser development effort was based upon a series of designs of experiments (DOEs). Jointly we and the laser manufacturer identified what we thought were the significant variables affecting the laser life. Thirteen variables were finally selected as the most promising to be evaluated, three of which were judged to be proprietary by the vendor. These sensitive process variables were associated with the oxidation process for the cathode and involved the atmospheric pressure, cathode current and time duration to oxidize the cathode. Their process levels were identified only by level 0.5, 1.0 or 1.5X. The first and most significant experiment tested ten variables at two levels and three more at three levels. This was accomplished in a Resolution IV experiment design of only 32 trials.[5] This test design is shown in Table I.

This initial study identified the most sensitive product parameters and their settings for building stable lasers. There was also a series of following experiments that better defined the interaction defects and refined the construction details. In the end the laser successfully met all of its application requirements, including the specified life of the machine, and this state-of-the-art component never caused a field problem in the printer. The effect of DOE on laser life was at least an order of magnitude improvement in life and stability, and brought a common focus to the challenge of increasing the laser's life. The DOE approach is an efficient test procedure that defends against unwanted variation in the experiment. The development of the test plan promotes better understanding of the problem by synergistically collecting the knowledge and insights of all those involved. This process can build teamwork in carrying out the experiment and analysing the results. The higher the involvement, the more the participants feel ownership and 'buy-in' to implementing the results.

The chipped weld problem

The situation under discussion relates to a timing chain manufacturing process. The particular chain used in this study had 58 links in it. This company had made timing chains for over 75 years. Production was periodically plagued by a faulty weld on one of its links. The problem would rise up for a while, the company engineers would work on the problem, and it would go away. In due course the problem would manifest itself again.

The defective weld or chipped pin problem is illustrated in Figure 4. When defective welds were found, the entire chain was discarded as scrap. The chipped pins were almost all found in visual inspection of the completed chain.

A consulting statistician met with the engineering team of the manufacturing company. The conventional wisdom was they felt a particular variable, called variable 'C' (for proprietary reasons), was causing the problem. This was the chief variable they adjusted every time the problem occurred, and the problem always went away—only to occur again at some later time.

So the statistician conducted a brainstorming session with the company engineers. The goal was to have everyone associated with the manufacturing process involved. This involvement ensured that all aspects of the process were considered. This involvement helped to create a common vision of the problem. This teamwork process promotes better 'buy-in' to the experiment to be run as well as implementing its results.

Twenty-two factors were identified as potentially significant to this problem. This of course included variable C and twenty-one other variables. The next step was to rank order the variables by their perceived importance. The number of variables considered most significant in this problem was finally narrowed to six. This was done by a vote of all the knowledgeable people meeting.

The experiment to be run was to be represented by a Hadamard matrix at a Resolution IV level. Each variable was tested at two levels. This is shown in Table II. All variables are to be run at either their high level (+) or their low level (−). The limits of the variables were at their process limits. It was known that when the variables were at some point in their design limits, good product ensued. The challenge was to understand where that good design point lies.

The experiment was conducted over a 12 month

Table I. Experimental design for 2^{10} 3^3 laser experiment

No. of trials NT	Getter A	Cathode material B	Wire-brushed or etched C	Cleaning D	Air firing E	Bake-out F	Geometry G	He:Ne ratio H	Fill pressure I	No. of cycles J	O_2 Pressure K	O_2 Current L	O_2 Time M
1	BULK	2024	WB	HEDD	900	300	MOD	20:1:1	2.15	4	0.5P	0.5C	1.5T
2	BULK	6061	WB	HEDD	0	300	MOD	17:1:1	2.50	2	0.5P	1.5C	1.0T
3	BULK	6061	E	HEDD	0	0	MOD	20:1:1	2.15	4	1.0P	1.0C	0.5T
4	BULK	6061	E	SPEC	0	300	PRES	20:1:1	2.50	2	0.5P	0.5C	1.0T
5	BULK	6061	E	SPEC	900	300	MOD	20:1:1	2.50	4	1.5P	1.5C	1.5T
6	COMB	6061	E	SPEC	900	0	MOD	17:1:1	2.50	4	0.5P	0.5C	1.0T
7	COMB	2024	E	SPEC	900	0	PRES	20:1:1	2.15	4	1.5P	0.5C	1.0T
8	BULK	2024	WB	SPEC	900	0	PRES	20:1:1	2.50	2	1.0P	1.0C	1.0T
9	BULK	6061	WB	HEDD	900	300	PRES	17:1:1	2.50	4	1.5P	0.5C	0.5T
10	COMB	6061	E	HEDD	0	300	MOD	17:1:1	2.15	4	1.0P	1.0C	0.5T
11	BULK	2024	E	SPEC	0	300	MOD	17:1:1	2.15	4	1.5P	0.5C	1.0T
12	COMB	6061	WB	SPEC	900	300	MOD	20:1:1	2.15	2	1.0P	1.0C	0.5T
13	COMB	2024	E	HEDD	900	300	MOD	20:1:1	2.50	2	1.0P	1.0C	0.5T
14	BULK	2024	WB	SPEC	0	0	MOD	20:1:1	2.50	4	1.0P	1.0C	0.5T
15	COMB	6061	WB	HEDD	900	0	PRES	20:1:1	2.50	4	0.5P	1.5C	1.0T
16	COMB	2024	E	HEDD	0	300	PRES	20:1:1	2.50	4	1.0P	1.0C	1.0T
17	COMB	2024	WB	SPEC	0	300	MOD	17:1:1	2.50	4	1.0P	1.0C	1.0T
18	COMB	2024	WB	HEDD	900	0	MOD	17:1:1	2.15	4	1.5P	1.5C	1.0T
19	BULK	2024	WB	HEDD	0	300	PRES	20:1:1	2.15	2	1.5P	1.5C	1.0T
20	COMB	6061	WB	HEDD	0	0	MOD	20:1:1	2.50	2	1.5P	0.5C	1.5T
21	BULK	2024	E	HEDD	0	0	PRES	17:1:1	2.50	4	1.0P	1.0C	1.5T
22	COMB	6061	WB	SPEC	0	300	PRES	20:1:1	2.15	4	1.0P	1.0C	1.0T
23	BULK	2024	E	HEDD	900	0	MOD	17:1:1	2.50	2	1.0P	1.0C	1.0T
24	BULK	6061	WB	SPEC	0	0	PRES	17:1:1	2.15	4	1.0P	1.0C	1.5T
25	BULK	6061	E	HEDD	900	0	PRES	20:1:1	2.15	2	1.0P	1.0C	1.0T
26	COMB	6061	E	SPEC	0	0	PRES	17:1:1	2.50	2	1.5P	1.5C	0.5T
27	BULK	2024	E	SPEC	900	300	PRES	17:1:1	2.15	4	0.5P	1.5C	0.5T
28	BULK	6061	WB	SPEC	900	0	MOD	17:1:1	2.15	2	1.0P	1.0C	1.0T
29	COMB	6061	E	HEDD	900	300	PRES	17:1:1	2.15	2	1.0P	1.0C	1.5T
30	COMB	2024	E	SPEC	0	0	MOD	20:1:1	2.15	2	0.5P	1.5C	1.5T
31	COMB	2024	WB	SPEC	900	300	PRES	17:1:1	2.50	2	1.0P	1.0C	1.5T
32	COMB	2024	WB	HEDD	0	0	PRES	17:1:1	2.15	2	0.5P	0.5C	0.5T

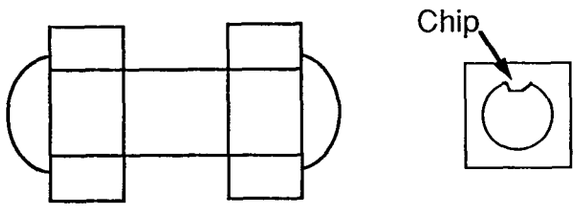

Figure 4. Design of experiments (pin chipping problem)

period as pieces were assembled from parts at their specified limits for each particular trial.

The results of the trials are shown in Table III with the results sorted top to bottom, best to worst.

The most discriminating variable is shown to be 'D'. When D is at its high level, product yields are good. These results are being depicted on what is termed an 'analysis of goodness (ANOG)' table. The effects of the experiment can be analysed graphically. No direct statistics need be used, so the results can be readily interpreted and believed by all members of the team.

This data is further refined by looking at the trial results for when D is at its high level. This is shown in ANOG Table IV. Here one can see the combined influences of other variables. Notably the best combination is high D, low A and high C. The dominant factor is still D.

TEST CONCERNS

One of the proverbial test concerns is whether to test the early prototypes while the design is still very fluid. Here the design changes may make the applicability of the test results obsolete. The item under test may no longer reflect the real design. Alternatively, one can wait and test the production-ready prototype. The problem then is that it is so difficult and expensive to change the design to correct any test problems found. In essence the test is too late. One solution to this problem is offered here. That is to perform early exploratory testing to find basic design limitations of the early designs. Problems found at this stage can quickly be fed back to the designers and accounted for in the design improvements that are typically ongoing at this stage of development. Once the design has matured and stabilized, it is more appropriate to do conventional life testing. The reliability and life data then obtained is indicative of the unit's performance in

Table II. Sixteen-trial Hadamard matrix

	Column of variables															
Run No.	0	A 1	B 2	C 3	D 4	5	6	7	E 8	9	10	F 11	G 12	13	H 14	15
1	+	+	−	−	−	+	−	−	+	+	−	+	−	+	+	+
2	+	+	+	−	−	−	+	−	−	+	+	−	+	−	+	+
3	+	+	+	+	−	−	−	+	−	−	+	+	−	+	−	+
4	+	+	+	+	+	−	−	−	+	−	−	+	+	−	+	−
5	+	−	+	+	+	+	−	−	−	+	−	−	+	+	−	+
6	+	+	−	+	+	+	+	−	−	−	+	−	−	+	+	−
7	+	−	+	−	+	+	+	+	−	−	−	+	−	−	+	+
8	+	+	−	+	−	+	+	+	+	−	−	−	+	−	−	+
9	+	+	+	−	+	−	+	+	+	+	−	−	−	+	−	−
10	+	−	+	+	−	+	−	+	+	+	+	−	−	−	+	−
11	+	−	−	+	+	−	+	−	+	+	+	+	−	−	−	+
12	+	+	−	−	+	+	−	+	−	+	+	+	+	−	−	−
	+	−	+	−	−	+	+	−	+	−	+	+	+	+	−	−
	+	−	−	+	−	−	+	+	−	+	−	+	+	+	+	−
15	+	−	−	−	+	−	−	+	+	−	+	−	+	+	+	+
16	+	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
alias	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
structure		A	B	C	D	AB	BC	CD		AC	BD	*	*	CE	*	AD
						DE	AF	EF	E	BF	AE	*	*	DF	*	BE
6 variables						CF						F	*		*	
7 variables							DH	BH		EG	CG		G	AG	*	FG
8 variables						GH	EH	AH		DH	FH			BH	H	CH

(Note: all interactions shown above are negative)
Resolution IV designs: 6 variables: A,B, C, D, E, F; 7 variables: A, B, C, D, E, F, G; 8 variables: A, B, C, D, E, F, G, H.
Main effects and interactions. Main effects clear (not aliased with 2-factor interactions: higher orders assumed negligible. Second order interactions aliased with each other. Error estimates from unlabelled columns (*).

Table III. Test results ranked by AOG

Trial	A	B	C	D	E	F	n
7	−	+	−	+	−	+	0
11	−	−	+	+	+	+	0
5	−	+	+	+	−	−	1
6	+	−	+	+	−	−	2
4	+	+	+	+	+	+	3
15	−	−	−	+	+	−	3
14	−	−	+	−	−	+	8
12	+	−	−	+	−	+	11
16	−	−	−	−	−	−	15
9	+	+	−	+	+	−	29
8	+	−	+	−	+	−	64
10	−	+	+	−	+	−	99
13	−	+	−	−	+	+	293
1	+	−	−	−	+	+	671
3	+	+	+	−	−	+	997
2	+	+	−	−	+	+	1065

Table IV. High level of 'D' pin chipping problem

Trial	A	B	C	D	E	F	*n*
7	−	+	−		−	+	0
11	−	−	+	+	+	+	0
5	−	+	+	+	−	−	1
6	+	−	+	+	−	−	2
4	+	+	+	+	+	+	3
15	−	−	−	+	+	−	3
12	+	−	−	+	−	+	11
9	+	+	−	+	+	−	39

its targeted application. This is true so long as the part is tested in a manner that simulates its use in the field.

Testing can be 40 per cent of the total development effort on a program. It can be the most expensive element in the development process. Testing has to be done on a good foundation or the results can be confused and misleading. The making of a good test foundation is the emphasis of this paper.

Now, a last look at some lessons learned to avoid common test pitfalls.

Lessons learned

In every programme that are some blunders that we would hate to repeat. These are often labelled 'Lessons learned' and typically are derived from programme reviews. These lessons form the basis of some guidelines that can be prescribed for good test practices. Some of these resulting guidelines for good test practice are enumerated here:

1. *Need to achieve confidence in your measurements* prior to exposing the device under

test (DUT) to environment or life testing. When possible this is best done by making measurements using more than one technique, that is using independent test techniques. For example you could weigh an object using a calibrated spring scale and a calibrated balance scale. Good confidence in the measurement results when the two independent measurement approaches agree with one another. The difference between the measurements could be labelled the 'error of closure'. This is the term used in sequential compass measurements. The analogy there is successive compass sightings are taken that lead one on a course where the sightings close back to the starting point. The closer the end point is to the starting point, the better the error of closure.

2. *Employ a gold measurement standard.* Set aside one or more samples that are preserved in their original state. These samples do not go through any stress testing. They are kept in a like-new condition. Every time that measurements are made on the samples under test, these gold standards are measured. Consistent measurement results on these units bolster confidence that the instrumentation has been stable and that the readings are largely repeatable. The early miners kept a canary in the mines with them. If noxious gas formed, it would kill the canary first. This signalled a gas build-up problem in the mine. Likewise, changing measurements of the gold standard signal that something has most probably changed in the measurement apparatus and subsequent measurements are suspect until resolved.

In software testing the gold standards are the regression test cases. These regression test cases are the standards representing the basic application of the code. As changes are made to the code to fix a bug, enhance the functionality, or adapt the code to a different base, test cases are run against the basic regression code to make sure that it still performs satisfactorily on its base problem set. Introducing changes to code often introduces a new error condition. The regression test cases attempt to verify that this is not happening and that the problem fix did not contaminate the code.

3. *What you see is not always what you think it is.* In one instance, one of the authors was testing a motor driver card that was experiencing excessive failures during manufacturing. The technician who was helping measured an excessive voltage from gate to cathode on a SCR. The specification was only a few volts, whereas the scope was showing hundreds of volts across this junction. The measurement was being made with a pair of differential scope probes. One was set on the gate and the other was placed on the cathode. Further investigation revealed that the probes were not sufficiently calibrated. This was shown by attempting to read the same pulse train with both probes. This differential reading should have been null since both probes were measuring the same waveform. Surprisingly, the differential probes were showing a 50 volt spike at each transition, which was totally false. The transient behaviour of these probes was far out of calibration. These probes read correctly when they had to measure steady-state conditions. However, we were looking for transient conditions that could drive the SCR into a second breakdown condition. Once the probes were set on the same calibration bar on the scope they could be easily adjusted so that would null out even under the transient condition. Note: this did not prove that the probes were now measuring correctly, but it did demonstrate that the probes were measuring incorrectly initially—even though you could clearly see the measurement on the scope screen. The final step was to have the probes recalibrated so one could have true confidence in their readings.

4. *Randomize uncontrolled variables.* One of the authors was running a comparison test of two manufacturers of solid tantalum capacitors. These units were powered at their rated voltage and put into an oven. The temperature was increased in a step-stress condition. The oven temperature was increased in 5 °C steps each day. Finally one of the vendors' units began to fail while the other kept running. This looked like tell-tale evidence that the one vendors' product was superior. The *confounding* problem later discovered was that there was a 15 °C gradient across the oven with the hotter temperatures being on the upper shelf where the parts were failing. The difference between the two vendors may have been in their respective locations in the oven, rather than in their quality. If their test positions in the oven and on the two shelves had been randomly assigned, this unwarranted confounding would have been checked. So randomizing uncontrolled variables neutralizes their possible effect on the test. There is a negative to such randomization however. Even though the confounding of effects is removed, the resultant uncontrolled variation in the oven increases the random error of the experiment. This makes the analysis harder by increasing the noise of the experiment. This noise makes it more difficult to see the true effects that the experiment is investigating.

Let us examining 'confounding of data' more closely, since it is a fundamental testing concern. Usually the testing goal is explicit, such as demonstrating an acceptable reliability level over a required life period. The experimenter is searching for causality. That is, the experimenter is attempting to correlate the change in the output variable with a change in one of the input or collateral variables. The variables are said to be confounded when the two of them are varying together, but not necessarily due to causation. In the tantalum capacitor experiment the observed output variable was the increasing stress on the capacitors over time, whereas the unrecognized input variable was the shelf position

(and the concurrent higher temperature of the upper shelf over the lower one).

CONCLUSION

Good test practice can bring a lot to the development process. It can be first used to review the product specifications. This is to see if the requirements specified can actually be tested and verified. Most importantly the performance of the unit under test can be explored over its operating domain. This is best done with the structure of design of experiments. DOE also can bring together the various interested parties to the product development process. It is an equilibrator making all contributions welcome and at par with one another. It is an excellent example of the concurrent engineering process working. Such a test structure leads to the development of robust designs. Reliability testing, typically using Weibull analysis (and often stress acceleration) is also necessary to prove life expectations. Also suggested above are some practical techniques and test practice to ensure repeatable data. Good data allows one to draw good conclusions about the effects of the testing.

Testing specialists on a design team make a significant contribution. Statistical specialists can make a big impact, particularly through displaying design of experiment strategies during new product development. The chief concern is that the statistician lacks domain-specific knowledge. Therefore the statistician must seek and interact with development in such a way that knowledge readily flows both ways. The brainstorming activity that kicks off DOE is an excellent team-building activity. Maintaining the team focus on DOE promotes 'buy-in' to the solution. This process can lead to extraordinary product improvements.

REFERENCES

1. W. Nelson, *Accelerated Testing*, Wiley Interscience, New York, 1990.
2. Stian Lyderson, *Accelerated Testing*, Royal Norwegian Council for Scientific and Industrial Research, #STF75A86010, March 1983. Also available from the National Technical Information Service (USA) #PB87-145157.
3. R. B. Abernathy, *The New Weibull Handbook*, 1993. Available from the author at 536 Oyster Road, North Palm Beach, Florida 33408, U.S.A.
4. Fultons Findings, *WeibullSMITH* Software package, available from Fulton Findings, 1251 W. Sepulveda Blvd., #800, Torrance, California 90502, U.S.A.
5. W. J. Diamond, *Practical Experiment Design for Engineers and Scientists*, Von Nostrand Reinhold, New York, 1989, p. 390.

Authors biographies:

Samuel Keene is the Product Test Manager at Storage Technology Corporation. A recognized authority on reliability and quality metrics and management, he has been actively involved in reliability and quality at IBM, NASA and Storage Technology for the past 30 years. He holds a Doctorate in Operations Research from the University of Colorado. Dr. Keene is a Senior member of the IEEE and is the past president of the IEEE Reliability Society. He has also published over a hundred papers in the field of reliability

Paul Vattano is a staff mechanical engineer at Storage Technology Corporation in Louisville, Colorado. His work involves design and testing of electromechanical parts and assemblies. Mr. Vattano has a B.S.M.E. from the University of Illinois and an M.S.M.E. from Colorado State University. He is a licenced professional engineer in Colorado.