Back to Session Subjects

TESTING COMPLEX HELICOPTER WEAPON SYSTEMS IN THE 21st CENTURY: ON THE RIGHT TRACK?

Dr Peter J Smith, Merlin Programme, Lockheed Martin ASIC

Figure 1: Royal Naval Merlin Helicopter

Abstract

Since World War I, the flight-testing of aircraft has been accomplished in a traditional manner. The manufacturer would establish a flight envelope and demonstrate structural soundness. Any deficiencies would be fixed by modifying or redesigning the offending part or system. Ground and flight-testing would ascertain which military specifications were met and whether the aircraft was suitable to carry out the intended mission. Prototype or pre-production hardware was evaluated so that several iterations could be tested and fixed prior to freezing the design. This methodology produced effective aircraft weapon systems, but development was slow, costs were high, threats changed, and advances in technology often made the system under test obsolete. The Ministry of Defence (MoD) decided to evolve the system when the time to field new systems became excessive and, in some instances, the final product failed to meet original requirements. Development costs were not only becoming "budget-busters" but were often open-ended, for example, the contractor would test-and-declare, and any deficiencies would be corrected by the customer. New techniques to reduce the development cycle and the risk to the customer had to be developed and incorporated. Enter the Royal Navy's Merlin Helicopter, designed to replace the venerable Sea King in both an Anti-submarine Warfare (ASW) and Anti-surface Warfare (ASuW) roles.

In the initial development contract signed in 1984, the contractor was responsible for the basic airframe, while MoD procured the mission equipment. Each mission subsystem had a separate specification and would be purchased individually. As the sophistication of an integrated weapon system with over 1.7 million lines of software code was realised, the MoD decided to redefine the specification in terms of mission capability and seek a system prime contractor to ensure this new goal was reached. In 1991, IBM (later Loral and now Lockheed Martin) was selected as prime contractor to manage the entire integration of the mission system and guarantee the performance of the weapon system. The challenge was to prove all the flight-related specifications within the maximum number of operational flight hours allotted under the fixed-price contract.

This paper discusses a technique of proving statistical compliance that limits both the buyer's risk (probability that a test will pass in a system that does not truly meet specification) and seller's risk (probability that a test will fail in a system that truly exceeds specification) using a minimum number of data points. The technique is called Sequential Probability Ratio Testing (SPRT), which results in

early termination of testing if either the specification requirements are met by a wide margin or the performance is markedly worse than the specification value. Only with marginal performance does the testing require additional data to be collected (up to the truncation point, Nmax) until a pass or fail result is obtained. Thus the sample size is not fixed in advance, but is determined during the course of the test by criteria which depend on the observations as they occur. This approach to testing not only limits the risks, but also provides a more efficient use of the very scarce and expensive flight hours.

The performance of the Radar subsystem, tested under a variety of sea states, will be examined to show the difficulties of flight testing in environmental conditions that are not strictly controlled. Unclassified results from the operational testing during Spring 2000 at the Atlantic Underwater Test and Evaluation Centre (AUTEC) against increasingly evasive submarines in an open ocean environment will also be discussed.

1 Introduction

Lockheed Martin ASIC is the prime contractor for the new Royal Navy Merlin Helicopter Weapon System. Entering service in 1998, the final of the 44 aircraft will be delivered to the customer by early 2002. The Merlin weapon system, Figure 1, is a highly complex combination of a large, long-range (EH101) airframe and advanced avionics equipment, which have been successfully integrated to enable its key mission requirements of anti-submarine and anti-surface warfare to be met.

As part of its prime contractorship, Lockheed Martin has been responsible for the operational performance flight trials of the Merlin and the associated cost and schedule. These trials have taken place in many diverse locations such as the Hebrides in Scotland, Aberporth in Wales and the Atlantic Undersea Test and Evaluation Centre (AUTEC) in the Bahamas. These extensive flight trials are the ultimate proof of the success of the Merlin Weapon System. As with most programmes of this nature, the flight trials were performed at the final stages of the contract and tend to be an extremely expensive and time consuming activity.

In order to prove the operational system requirements satisfactorily to the customer, minimise costs, maintain schedule and effectively plan the flight trials programme, Lockheed Martin has used the SPRT statistical technique. This has paid significant dividends by saving flight hours and providing statistical proof that the system meets its operational requirements.

2 Traditional Approaches

Historically, flight test programmes have used a strict “single-line” approach to development. Manufacturers would establish a flight envelope and demonstrate structural soundness. Any deficiencies would be fixed by modifying or redesigning the offending part or system. Development would then be passed to ground and flight-testing to ascertain which military specifications were met and whether the aircraft was suitable to carry out the intended mission. Prototype or pre-production hardware was evaluated so several iterations could be tested and fixed prior to freezing the design. It was only after the start of production that operational forces were able to assess the system. This methodology produced effective aircraft weapon systems, but development was very slow, costs were high, threats changed, and advances in technology often made the system obsolete. The cost of fixing problems in the early stages of the programme were relatively inexpensive compared to the cost of fixing problems found when the operational forces tested the system. Moreover, the testing techniques were often of the “fixed sample” variety. Such testing resulted in a considerable time lapse between the completion of testing and the publication of the results which consequently delays decisions on required fixes. This added further to the costs of development.

Lockheed Martin (formally, IBM Federal Systems) challenged this traditional approach with a new philosophy, which was applied to the US LAMPS helicopter programme during the 1980’s. As prime contractor, a “full scale development” programme approach was employed that introduced more parallel activities, combined with early involvement of the military operators, and increased use of modelling and simulation techniques. Moreover, the programme moved away from fixed-sample size testing and employed SPRT techniques to bring even further cost and schedule savings.

Having been successfully used on the US LAMPS programme, these techniques were repeated when the UK’s Merlin Helicopter programme was awarded to Lockheed Martin in 1991. The SPRT philosophy was also applied to the Merlin operational trials. Thus, as the turn of the century approached, the traditional development cycle had been replaced with a more cost-effective, parallel, customer-involved methodology, whilst maintaining standards and producing a high quality product.

3 Sequential Testing

As a pre-cursor to discussing SPRT methodology and results from the Merlin flight trials it is worthwhile introducing the concept of sequential testing. Sequential testing techniques do not fix the sample size prior to testing, but offer the flexibility to terminate testing as soon as sufficient data is collected to provide a significant conclusion. In very simple terms, consider the tossing of a coin in order to test for bias. It would be perfectly acceptable to get two heads after two throws of the coin. However, if ten throws gave ten heads then there is probably sufficient evidence to point to a biased coin and for testing to stop.

Applying this concept further, suppose a random sample of 25 air filters were chosen from a production contract and it was agreed to reject the complete production batch if 3 or more of the sample were defective. Figure 2 illustrates a possible outcome. At sample number 13 the third failure has occurred and therefore testing can stop. This sequential method has saved testing filters 14 to 25, particularly significant if testing is time consuming and or expensive. Traditional fixed sampling methods would have tested all 25 units. More generally, it is possible to draw an accept line and a reject line where testing can stop if the test results reach either of these limits.

Thus, the use of sequential sampling enables the efficient reallocation of expensive test resources when a test demonstrates "early on" that the performance of a system is either significantly better or significantly worse than the given requirement.

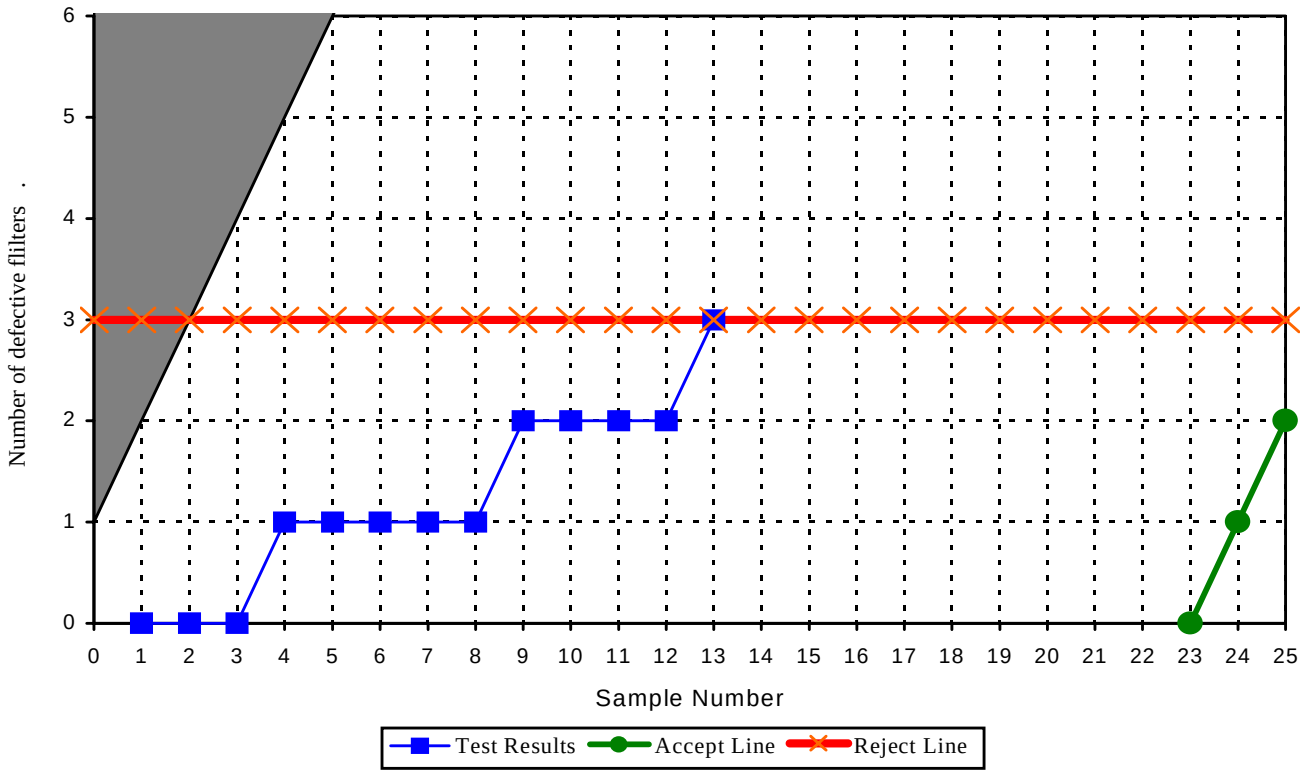
4 SPRT Methodology

SPRT takes this relatively straight forward sequential approach and applies it using advanced hypothesis testing and an assumed statistical distribution to produce pass/fail curves for a variety of performance measurements. The following have been used in the Merlin programme:

(a) Proportion defective (Binomial distribution) For example, proportion of radar targets not detected.

(b) Root Mean Square (RMS) error (One dimensional normal distribution). For example, accuracy of an aircraft altimeter.

(c) Circular Error Probability (CEP) (Two dimensional normal distribution). For example, accuracy of weapon drop on surface of the sea.



The SPRT methodology uses θ (theta) to represent the true (but unknown) performance parameter of the system under test. Moreover, θ_{spec} represents the specification performance requirement. Generally, large values of θ indicate poor system performance and therefore θ_{spec} is the maximum value of θ allowed by the system specification.

For example, if testing a radar's ability to detect a small target, the performance parameter, θ, could be the proportion of non-detections with the specification performance requirement, $\theta_{spec} = 0.1$. Another typical scenario could be testing the ability of an aircraft to deploy a weapon to hit a target (for example, delivery of a torpedo to within a certain radius of a target position on the surface of the sea). In this instance, the performance parameter could be the Circular Error Probability (CEP) with the specification performance requirement being say, θ_{spec} = 850 metres. (Note: CEP is the radius of a circle that contains 50% of the probability).

The SPRT technique tests the following hypothesis:

Null hypothesis $H_0 : \theta <= \theta_{spec}$
(Performance parameter is better than or equal to specification)

Alternative hypothesis $H_1 : \theta >= \lambda\theta_{spec}$
(Performance parameter is worse than or equal to λ times specification)

The constant λ (lambda) is called the discrimination ratio. It can be considered to be a tolerance window above the specification performance requirement. The value of λ is determined by negotiations between the producer (seller) and the customer (buyer) and generally falls within the range of 1.05 to 1.20. Values less than 1.05 require large quantities of test data, whereas large values do not inspire customer (buyer) confidence. (Note: λ is always greater than 1.0).

At any stage of a test (ie, the collecting of discrete test results) that uses SPRT a decision is made to:

(a) Accept the Null hypothesis (H_0) and stop testing (pass)
or (b) Accept the alternative hypothesis (H_1) and stop testing (fail)
or (c) Continue testing by taking an additional sample (continue).

A failure of a test initiates a search for the cause and related fixes. After fixes have been implemented, the trial is repeated. This process is illustrated in Figure 3.

This type of test can lead to two kinds of error the probabilities of which are denoted by α (seller's risk) and β (buyer's risk). The seller's risk (α) is defined as the upper limit on the probability that a test will fail a system that truly meets or is better than specification. The buyer's risk (β) is the upper limit on the probability that a test will pass a system that is truly worse than lambda times specification. Clearly, the buyer and seller risks should be as small as possible.

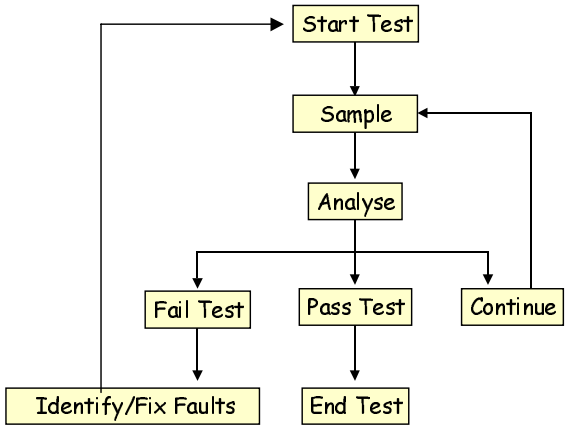


Figure 3 – General Process for Sequential Sampling

Before testing commences, α, β and λ are all agreed and the pass/fail curve is produced. The negotiations to reach these agreements depend on the allocation of test resources (eg flight hours). The resources reflect programme priority and can be quantitatively adjusted using α, β and λ. In the case of the Merlin programme, these parameters were agreed with the customer between 1992 and 1995 for flight trials that took place between 1998 and 2000.

The pass and fail curves meet at infinity (at the centre line value). It is therefore possible, where the system performance is close to the specification value, for the test statistic to remain in the continuation region for a significant period and require a large amount of testing before a conclusion is reached. To remove the possibility of this unbounded testing, the Merlin programme used truncated SPRT tests where a maximum number of samples (denoted by N_{max}) was agreed prior to testing. At or by N_{max} samples, a pass or fail decision will always be made. If no pass/fail decision has been made before N_{max} samples, the test will pass at N_{max} if the test statistic is less than, or equal to, the centre-line value and fail if the test statistic is greater than the centre-line value.

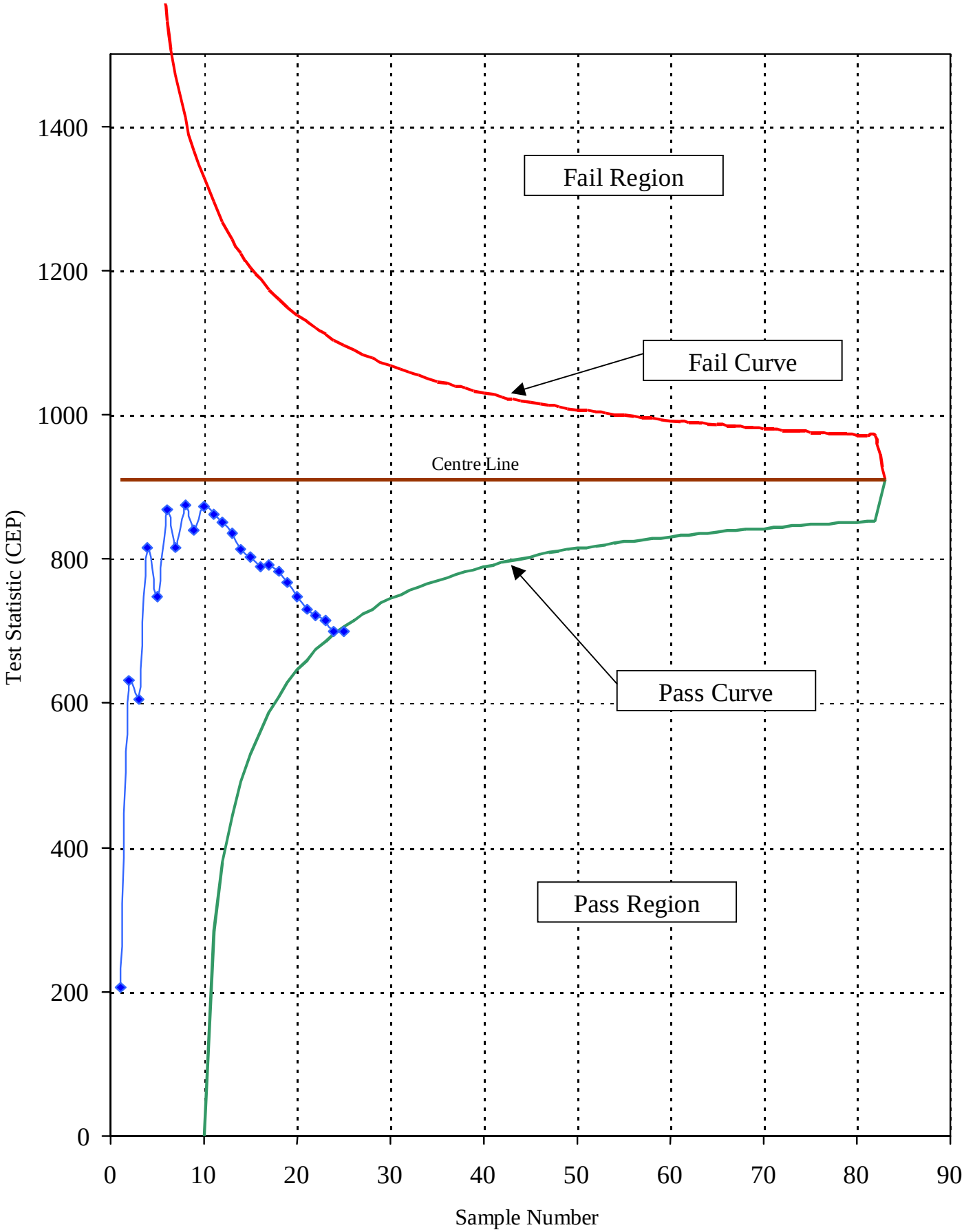


Figure 4 - Example SPRT Pass/Fail curves

Sellers Risk (α) = 11% Buyers Risk (β) = 9% Discrimination Ratio (λ) = 1.15

Circular Error Probability (CEP) using a 2D Normal Model

Limiting the maximum number of samples increases the buyer and seller risks.

Figure 4 illustrates an example of pass/fail curves for a two-dimensional normal model with seller's risk (α) of 11%, buyer's risk (β) of 9%, discrimination ratio (λ) of 1.15 and specification performance requirement (θ_{spec}) = 850 (CEP) metres. The three (pass, fail and continue) regions are defined by the pass/fail lines as shown. It is on this graph that the test statistic is plotted as each sample data point is collected until a pass or fail decision is reached. An example test statistic is plotted, which shows a pass at sample number 25.

5 SPRT Case History 1 - Radar Small Target Detection

Of all the Merlin flight trials that have used SPRT, only one, to date, namely radar small target detection, produced an initial fail and necessitated some changes to the operational techniques used to conduct the trial. The Merlin radar is required to detect reasonably small targets at sea from reasonably long ranges (actual figures are not given here for reasons of security classification). Such results are highly dependent on the sea-state. The rougher the sea the harder it is to locate targets, particularly small targets. The trials were required to be conducted between sea-states of 2 to 4.

In December 1998, the Merlin flight trials team deployed to Aberporth in Wales to prove this radar small target system requirement. This binomial SPRT (that is, either the small target was detected or it wasn't) produced an overall fail result after only 27 data points had been collected over a very few sorties (see figure 5a). At this point the trial was stopped and despite the unwelcome overall result, precious flight hours were saved because the system was performing so poorly. The Nmax value for this trial was 135, thus only 20% of the potential sample size was collected in order to reach a valid conclusion. A "get well" programme was initiated in order to identify and fix the underlying cause of the failure. The data collected during December 1998 was between sea-states 2 to 4. In simple terms a sea-state 2 equates to an average wave height of 1 to 2 feet and a sea-state 4 equates to an average wave height of 4 feet.

In July 1999, following a successful de-risk flight at relatively low sea state, the trial was repeated. Various improvements had been made to the trials operational procedures and one of the six small-targets, which was poorly located, was removed. The first sortie on day one gave promising results with only one failure in 21 (4.8%) samples collected, see figure 5b. The sea-state was just above 2. The following day the weather had deteriorated and with a sea state of around 3 two further sorties were flown. The results from these sorties were less promising (with 33.3% failures) and took the test statistic curve further from the pass line. On the third day the sea state had increased to just short of 4 and the remaining two sorties gave an even higher number of failures (42.9% and 57.1% respectively) resulting in an overall fail after 95 data points (out of a possible 135) at which point the trial was terminated. More detailed investigation and analysis of both trials followed and it was found that the following reasons were the primary cause of failure :

(a) The radar small target detection was more dramatically affected by sea state than originally expected. Low sea state data from the de-risk flight and the first trials sortie led Lockheed Martin to expect the system performance to be better than actual.

(b) Radar clutter characteristics. The motion and breaking of the sea is shown on radar as clutter. The method for suppressing clutter (via gain control) required modification in order to reduce the effect of "unstable" clutter and therefore cause the "stable" target to be more easily identified.

(c) Scan-to-scan Integration. The radar contained some scan-to-scan processing whereby a radar return in a similar position would show more prominently on the radar display than "random" sea surface activity after repeated scans over the same area. This function was not used during the first two Aberporth trials.

Finally, in November 1999, after these radar modifications had been implemented, the Merlin returned to Aberporth and produced a SPRT pass result after only 41 data points had been collected in a representative sample of sea states, see Figure 5c. The trial had finally produced a pass result and despite the problems identified (and fixed), only a minimum number of flight hours had been expended.

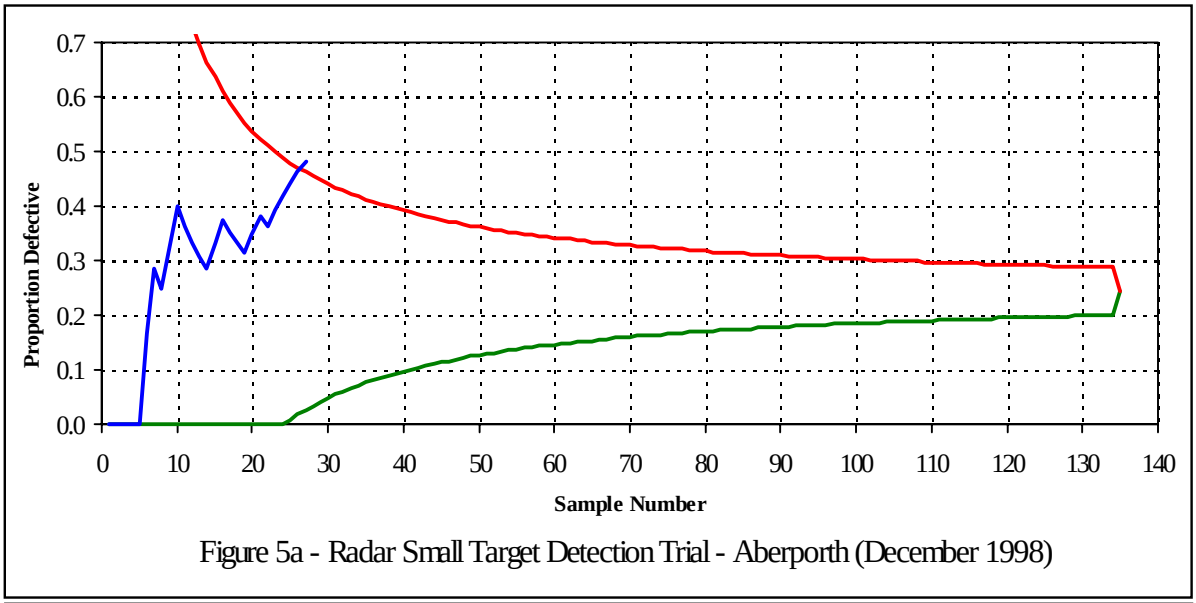


Figure 5a - Radar Small Target Detection Trial - Aberporth (December 1998)

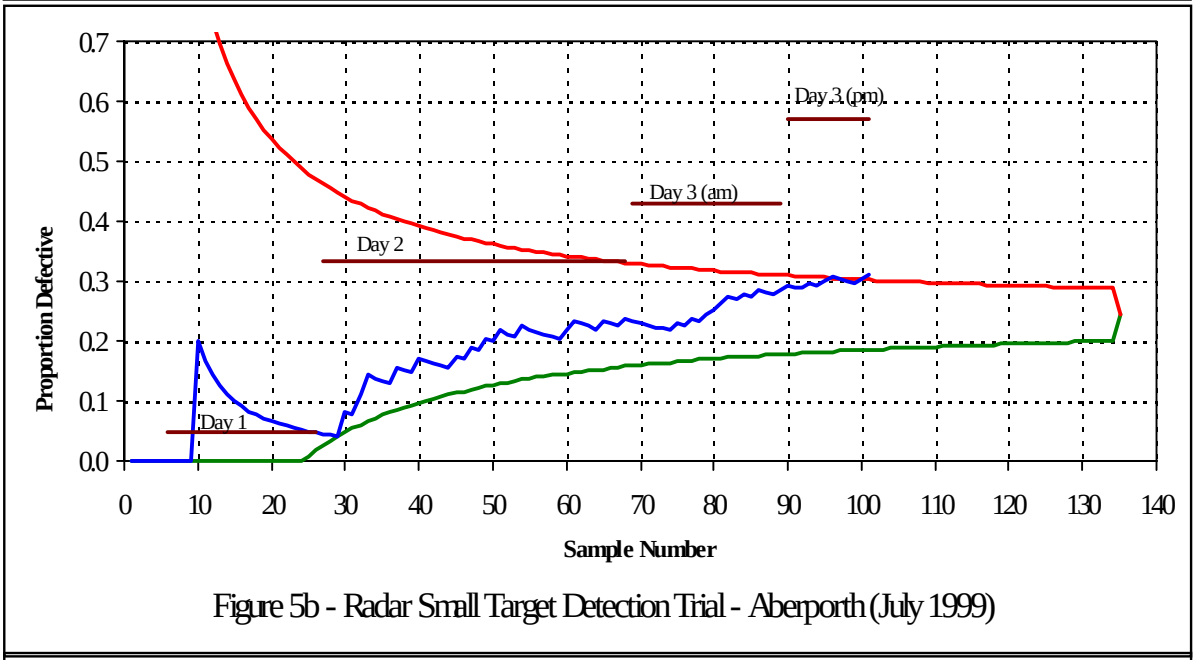


Figure 5b - Radar Small Target Detection Trial - Aberporth (July 1999)

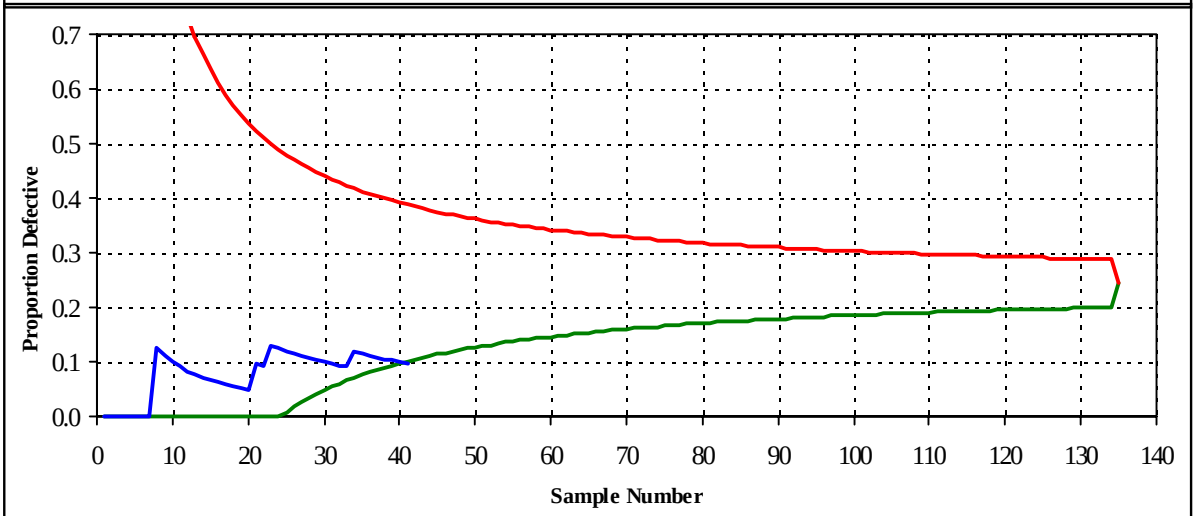


Figure 5c - Radar Small Target Detection Trial - Aberporth (December 1999)

This trial illustrates the SPRT process. It enabled a poor performing system to be identified early without wasting valuable flight hours. Further testing was only carried out after changes had been made. Although more dramatic changes should have been made after the first trial, the final deployment very quickly proved that the system now exceed the system operational requirement by passing the trial early.

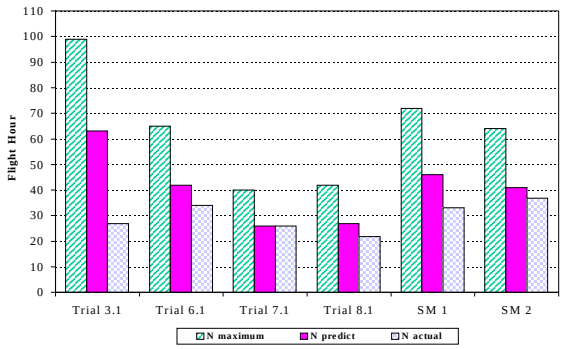


Figure 6: AUTEC Flight Hours

6 SPRT Case History 2 - Localisation and Prosecution of Submarines

In spring 2000, Lockheed Martin led an extensive period of Merlin flight trials at AUTEC, Andros Island, Bahamas. These Anti-submarine trials comprised six major mission scenarios and associated system requirements and were to be proved using ten SPRT tests. Using the SPRT methodology, with agreed parameters, it had been determined that a maximum of 382 flight hours would be required to complete all six AUTEC trials, assuming each test required N_{max} samples.

However, prior to the trials, the performance of the system had been estimated using modelling and development experience and it was predicted that, by exiting testing earlier than the N_{max} value, that all of the trials would require less data points and the total number of flight hours was predicted to be 245.

During the execution of the trials, it became apparent that the Merlin weapon system was far exceeding even the most optimistic estimates of the specification requirements and as a result, far less samples (and therefore flight hours) were actually required. In fact the Merlin exceeded requirements and every one of the ten SPRT tests gave early pass results. The total number of hours used during the deployment was 177 hours, see Table 1 and Figure 6.

	Flight Hours		
Trial	Nmax	Predicted	Actual
3.1 – Sonobuoy Localisation	99	63	27
6.1 – ADS Localisation	65	42	33
7.1 – Vectored Attacks	40	26	25
8.1 – Infotac	42	27	22
SM1 – Stressing Mission 1	72	46	32
SM2 – Stressing Mission 2	64	41	37
Total	382	245	176

Table 1: AUTEC Flight Hours – Spring 2000

Using sequential testing during these trials required a highly level of flexibility of those involved. After each flight, the data was analysed and entered into SPRT, usually within a few hours of the aircraft landing. It was then possible to determine whether the trial required further flights. As a result, it was difficult to plan trials activities rigidly. Those involved successfully managed the programme on a day-to-day basis by providing the flexibility to change the trials to be flown at relatively short notice.

The successful performance of the Merlin, combined with the use of SPRT, had freed additional flight hours, which were made available to the customer to perform specific sorties and trials to further test the Merlin's capabilities using the instrumented ranges a full year in advance of the first Royal Navy deployment in AUTEC.

7 Conclusion

This paper has provided a brief overview of the methodology and application of SPRT to the testing of the Merlin weapon system. The tool has proved its value in the testing of operational requirements in real-life scenarios with real-life targets. The case histories have shown that in all cases to date that the Merlin now exceeds its requirements and this has resulted in an early exit of testing and significant savings in flight hours. In situations where testing is time-consuming and/or expensive, this innovative technique provides valuable savings to all involved. As government contracts move more towards "Integrated Project Teams" and "Better, Faster, Cheaper" methodologies, SPRT proves to be an ideal tool by :

- Assisting the planning process by having a pre-determined maximum number (N_{max}) of samples for each trial.

- Providing flexibility for trials. Trials that pass early provide "spare" flight hours to support the re-testing of trials that fail.
- Systems that require fixing are better understood and usually show significant improvement on subsequent testing resulting in a pass with a small number of datapoints.
- In planning stages, SPRT provides a way to quantitatively discuss customer priorities and by adjusting λ, α and β, assign resources in a way that reflect customer priorities.

As the end of the Merlin operational trials approach it has become clear that the Merlin weapon system has not only met and exceeded it's operational requirements, but has also delighted the customer set. As the Royal Navy's Commanding Officer of the Merlin 824 Squadron recently reported "We are thrilled to bits with the aircraft and believe it is truly an aircraft for the 21st century"

8 References

1. SMITH, P.J. *Weapon Systems Testing: Knowing when to stop*; RAeS conference, Feb 1998
2. WALD, A. *Sequential Analysis*. New York: Dover Publications, 1947.
3. WALD, A. *Sequential Tests of Statistical Hypothesis*. Annals of Mathematical Statistics. 1945. Vol16, pp117-186
4. PITMAN, G. *Inertial Guidance*. John Wiley and Sons, 1962
5. ANDERSON, T.W. *A Modification of the Sequential Probability Ratio Test to Reduce the Sample Size*. Ann.Math.Stat. 1959. Vol. 31, pp 165-197.
6. WETHERILL, G.B and GLAZEBROOK,K.D. *Sequential Methods in Statistics*. Chapman and Hall, 1986.
7. MONTGOMERY, D.C and RUNGER, G.C. *Applied Statistics and Probability for Engineers*. John Wiley and Sons, 1994.