Annex B – MEASURES OF PERFORMANCE (MOPs)

B.1 INTRODUCTION

This Handbook focuses on EW Developmental Test & Evaluation **(**DT&E) and consequently MOPs are the central metrics. It is important, however to understand how MOPs fit into the overall hierarchy of test requirements, objectives, and associated measures. It is also important to understand what a measurement is and what information it conveys. Finally, this annex discusses some common MOPs. It is not intended to be definitive or an exhaustive compilation. It is intended to make the reader think about what details need to be addressed and documented in the planning stages to avoid disagreements later in the programme when they are much more difficult to resolve.

B.2 REQUIREMENTS, OBJECTIVES, AND MEASURES

Test requirements ultimately derive from operational needs identified by the military end user. These requirements are expressed as Critical Operational Issues (COI) and are defined as: "A key Operational Effectiveness (OE) and/or Operational Suitability (OS) issue (not a parameter, objective, or threshold) that must be examined in Operational Test and Evaluation (OT&E) to determine the system's capability to perform its mission. A COI is normally phrased as a question that must be answered in order to properly evaluate OE (e.g., "Will the system detect the threat in a combat environment at adequate range to allow successful engagement?") or OS (e.g., "Will the system be safe to operate in a combat environment?"). A COI may be decomposed into a set of Measures Of Effectiveness (MOE) and/or Measures Of Performance (MOP), and Measures of Suitability (MOS)." [1] Furthermore, the MOE, MOP, and MOS are defined as:

- **MOE**: Measure designed to correspond to accomplishment of mission objectives and achievement of desired results. MOEs may be further decomposed into Measures of Performance and Measures of Suitability. [2]
- **MOP**: Measure of a system's performance expressed as speed, payload, range, time on station, frequency, or other distinctly quantifiable performance features. Several MOPs and/or Measures of Suitability may be related to the achievement of a particular Measure Of Effectiveness (MOE). [3]
- **MOS**: Measure of an item's ability to be supported in its intended operational environment. MOSs typically relate to readiness or operational availability, and hence reliability, maintainability, and the item's support structure. [4]

MOPs are most commonly encountered as contractual specification requirements or other DT&E requirements. Some examples include: response times, Angle Of Arrival (AOA) measurement error, maximum detection range, etc.

B.3 MEASUREMENTS

One of the most important axioms in T&E is that system requirements must be testable. This means that the test must produce a meaningful answer to the questions asked. Whether or not a system meets its requirements will usually be determined by a measurement or series of measurements.

Measurement theory and statistics are complex fields and detailed treatments are beyond the scope of this Handbook. A measurement, by one definition, "in the broadest sense, is defined as the assignment of numerals to objects or events according to rules." [5] While there is controversy among statisticians

regarding the four scales or classifications of measurement shown in Table B-1, they serve as a good starting point for a discussion of MOPs.

Table B-1: Measurement Scales [6].

Scale	Attributes	Permissible Statistics (Examples)	Common Examples
Nominal	Classification only.	Number of Cases, Mode	First Names
Ordinal	Rank ordered – the differences between the values are not meaningful.	Median, Percentile	Hardness of Minerals, Quality of Leather
Interval	Uses a scale with an arbitrary zero point (can have numbers less than zero) – differences between values are meaningful, but ratios of values are not meaningful, i.e., 60°F is not twice as "hot" as 30°F.	Mean, Standard Deviation, Correlation, Regression, Analysis of Variance	Fahrenheit or Celsius Temperature Scales
Ratio	Uses a scale with an non-arbitrary zero point (cannot have numbers less than zero) – ratios of values are meaningful, i.e., a weight of 20 lbs. is twice as much as 10 lbs.	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Rankin or Kelvin Temperature Scales

The individuals charged with generating specification requirements should consult with experienced testers and analysts. This ensures that types of measurements are appropriate to the task and that the required data can be collected in sufficient quantities and at sufficient rates. Proper consideration of the measurements and associated analysis techniques will not only help answer the question of whether or not a System Under Test (SUT) meets its specification requirements, but will also support a broader characterization of SUT performance.

Data analysts should strive to choose the measurement scale that retains the maximum amount of information. Information retention can be illustrated using bombing MOPs as an example. Consider a specification where a hit or miss is determined by a specified bomb miss distance (a nominal measurement). A significant amount of information is lost by just evaluating whether or not each bomb produces a hit or a miss. By focusing the analysis on vector miss distances (a ratio measurement); analysts can determine much about the system by analyzing the range and direction of the errors.

B.4 MOP CONSIDERATIONS

Test designers must consider MOPs in light of how a SUT functions. As an example, consider the objective of evaluating the performance of a Radar Warning Receiver (RWR). The evaluation needs to address several different MOPs. The specification requirements will be expressed as MOPs.

In a perfect world, the system contractual specification requirements document would define not only the specific MOPs to be evaluated but also:

- The specific conditions under which the data will be collected.
- The data reduction, analysis, evaluation process, including statistical treatments.

Failure to address these considerations can cause unexpected variability in the results and possibly incorrect measurements and inaccurate results. The simplified case shown Figure B-1 illustrates how this can occur. The dwell structure shown could have a significant effect on the ability of the RWR to detect and identify radar A. If radar A is operating between 9.0 and 10.0 GHz the RWR doesn't need to make any decisions when a signal is detected; the measured frequency and Pulse Repetition Interval (PRI) are sufficient to make a unique identification. If, however, radar A is operating at 8.5 GHz the Mission Data File (MDF) has two other radars with which to contend. Assume that Radar B is ambiguous in frequency and PRI with radar A then the RWR will need to do additional processing to resolve the ambiguity; perhaps by determining scan type or rate or by a more detailed pulse analysis.

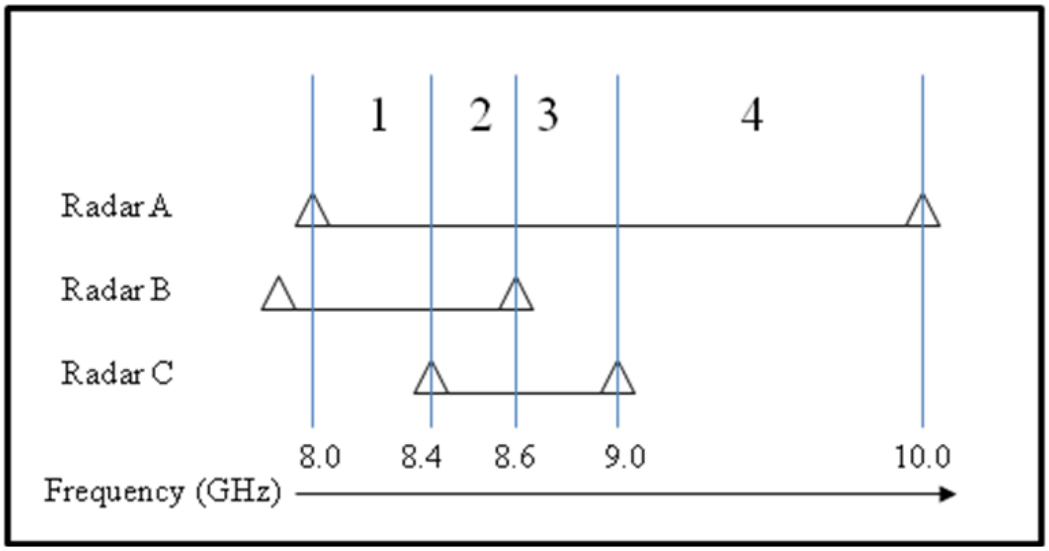


Figure B-1: Notional RWR Dwell Structure.

Therefore, it is likely that when radar A is operating at 8.5 GHz the response time will increase due to the additional possessing required to resolve the ambiguity with radar B and the likelihood of a misidentification also increases. If a test team elected to conduct a majority of the testing using radar A operating in frequency region 4, the RWR performance could be radically different than if it occurred in regions 1, 2, or 3. Test designers need to be aware of conditions such as this and consider them when designing test matrices.

B.5 SELECTED MOPs

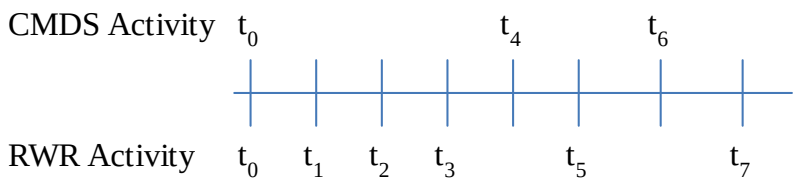
B.5.1 Receiver MOPs

This section addresses receiver MOPs by focusing on their applicability to RWRs, although they may also be applied to other receiver applications.

B.5.1.1 Response Time

Response time is one of the most important MOPs. It is a ratio measurement since there is an absolute zero reference. Generically, it is the elapsed time between two events. A federated or integrated EW suite may

have several response times associated with its performance. Figure B-2 shows some of the response times associated with the simple case of a federated EW system with an RWR serving as the EW data bus controller and a mission computer serving as the avionics data bus controller. In this case, the two most important response times are, from a military utility viewpoint, the time between the illumination by a hostile radar and the time that the system warns the aircrew and the time of the CMDS dispense. However, from a T&E and systems engineering standpoint each of the intermediate time intervals are also critical.



t_0 - Time of initial valid illumination by a radar
t_1 - Time of initial detection
t_2 - Time of initial identification
t_3 - Time display message made available on the avionics data bus
t_4 - Time that threat information is made available to the CMDS via EW data bus
t_5 - Time mission computer commands display
t_6 - Time of CMDS dispense
t_7 - Time of display to aircrew

Figure B-2: Response Times.

Contractual system specification requirements should clearly identify the system response time budgets that support the overall mission requirements such as the time to display a warning to the aircrew or generate a CMDS dispense. The simple case described above could involve up to three separate contractors: the RWR manufacturer, the CMDS manufacturer, and the airframe integration contractor.

When there is a deficiency in the overall system performance it is important to be able to identify the specific deficiency and who is responsible. The RWR manufacturer only controls the sequence of events leading up to making a display message available on the avionics data bus. The time between the message availability and the mission computer processing it, sending it to the display generator, and generating the display is under the control of the avionics integration contractor. Similarly, the CMDS manufacturer only controls the activity subsequent to receiving data bus messages.

Note that the response time MOP does not necessarily require a correct identification. Although not ideal, if the RWR displays an incorrect symbol and generates an audible warning tone in a timely manner the aircrew still has an opportunity to react. It is better to have an incorrect symbol displayed rapidly than to display the correct symbol when it is too late. If a system incorrectly identifies a threat radar it will be penalized using other MOPs such as percent correct identification.

Response time data, as with other data, can be described by the central tendency and the spread of the data. They are rarely normally distributed and usually skewed to the right. Each individual response time

must be greater than zero, while occasionally, the system will fail to generate a warning and the response time will be effectively infinite.

An average response time value must be treated carefully. First, consider using the mean. When a system fails to generate a warning, a mean cannot be calculated directly. One method of computing a mean when a data sample includes non-response is to transform the data.

Take a simple case with three samples: 2.0 second, 4.0 seconds, and no response. A mean value can be determined by transforming the data by taking the inverse of each value: 0.5, 0.25, and 0.0. The mean of the transformed data is 0.25. Transforming the data again produces a mean value of 4.0 seconds. Non-responses do not pose a problem for computing the median. In any case the test team must agree on the data analysis methods.

The skewed distribution poses a problem for evaluating the spread of the data. For this reason, RWR response time specifications are commonly expressed as percentiles, for example: 90% of the responses shall be less than X seconds. This method has the advantage of being easy to compute and gives some insight into both the central tendency and the spread of the data. Figure B-3 shows a hypothetical data set for a threat system with an acquisition radar, a target tracking radar, and a missile guidance radar; each with its own specification requirement. The box and whiskers plot is an effective way of presenting the data. In Figure B-3, if the response time specifications are 90% less than X seconds, then the RWR meets specifications for the missile guidance and acquisition radars, but does not meet specification for the target tracking radar.

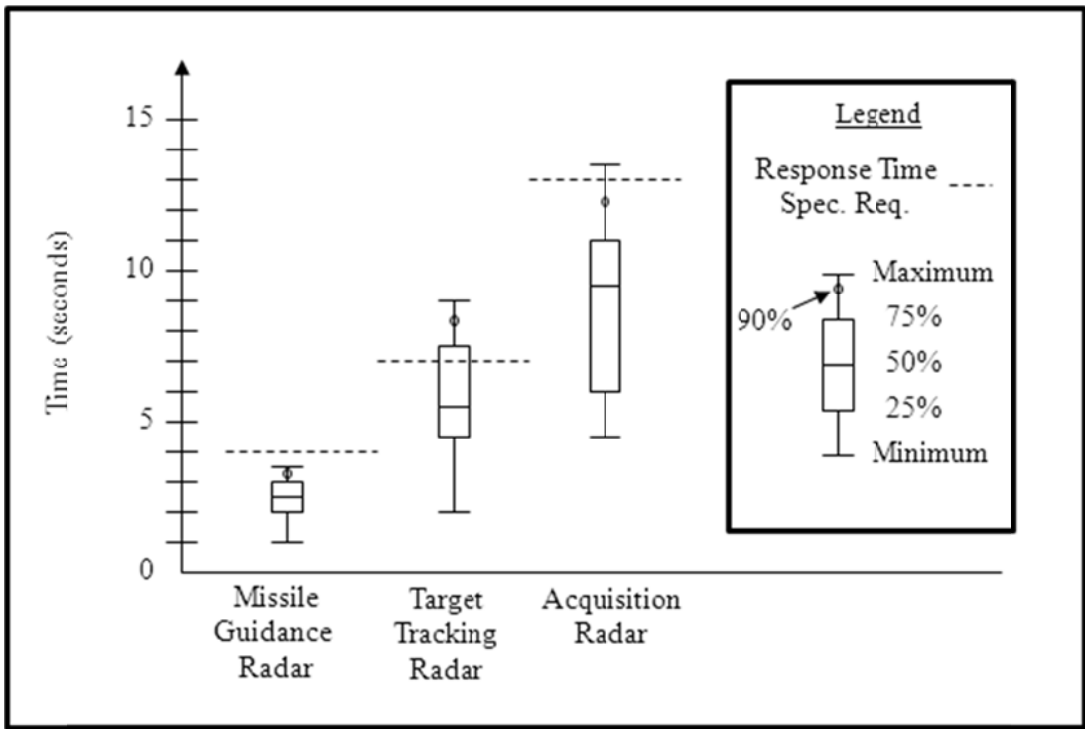


Figure B-3: Percentile Specification for a Hypothetical RWR (Not Real Data).

B.5.1.2 Correct Initial Identification Percentage

Radar directed threat systems are often composed of multiple beams. Each time an RWR is presented with a radar beam the system has an opportunity to correctly identify the beam. The correct identification

percentage is simply the ratio of the number of correct identifications to the number of identification opportunities multiplied by 100.

B.5.1.3 Correct Beam Correlation Percentage

When multiple beams are present the RWR should internally identify each beam individually and correlate them such that only a single symbol associated with the most lethal operating condition of the threat radar is displayed. Anytime more than one beam is present the system has an opportunity to correctly correlate them or to not correlate them if they are from different radar systems. The correct correlation percentage is simply the ratio of the number of correct beam correlations to the number of correlation opportunities multiplied by 100.

B.5.1.4 Correct Mode Change Percentage

Some radars, such as airborne fire control radars, only have a single beam with which to perform multiple functions. A common engagement sequence of events would be for the radar to transition from a search mode, to target tracking mode, and ultimately to a missile launch mode. An RWR should detect each mode change by the radar and update its track files. Every time that a radar transitions modes there is an opportunity for the RWR to correctly detect and process the change. The correct mode change percentage is simply the ratio of the number of correct mode changes to the number of mode change opportunities multiplied by 100. Each mode transition also presents opportunities to collect response time and identification data.

B.5.1.5 Maximum Detection Range

Receivers are typically required to detect specific signals at a specified maximum detection ranges. The measure can be accomplished in flight but it is time consuming to collect enough data to support a statistically meaningful assessment. This is particularly true in the case when a scanning receiver is attempting to detect a scanning radar. Hence, maximum detection range is a measure best evaluated analytically. In reality, the maximum detection range will be described by a statistical distribution. The power density associated with a given signal at the maximum detection range can easily be calculated and compared with the installed sensitivity to determine if the signal will be detected at that range. The installed sensitivity of a receiver is the product of the receiver sensitivity, transmission line losses, amplifier gain (if present in the installation), and the antenna gain and can be easily calculated. The receiver sensitivity and amplifier gain can be measured in a laboratory, the RF transmission line losses can be measured on the aircraft, and the antenna gain patterns can be obtained.

B.5.1.6 Angle Of Arrival (AOA) Measurement Accuracy

EW receiver systems have widely varying AOA accuracy requirements and depend on the purpose of the system, although most specify angular fields of regard. RWRs typically specify a 360 degree azimuth field of regard and are bounded by elevation bands.

The AOA accuracy is determined by analyzing the AOA measurement errors, where AOA error is defined as the difference between the AOA calculated by the system and the true AOA. The error data are often presented by angular bins; commonly as a Root-Mean-Squared (RMS) error versus angular bins.

AOA data can be complicated to analyze. If each measurement is an independent sample the analysis is relatively straightforward. However, most EW systems don’t present raw AOA measurements; they typically filter or smooth the data before it is presented or applied. When AOA data are filtered or smoothed, they are no longer independent and care must be taken when performing statistical analyses. Professional statisticians should be consulted when dealing with non-independent data.

B.5.1.7 Geolocation Accuracy

Some EW systems need to determine the location of ground-based emitters. These systems commonly measure the AOA to the emitter and based on successive measurements and triangulation algorithms produce an error ellipse that should contain the emitter's location as shown in Figure B-4. The speed with which a system can accurately locate an emitter is a function of geometry; it takes longer to locate an emitter off the nose of the aircraft than one off the beam due to the less rapid change in absolute bearing to the emitter.

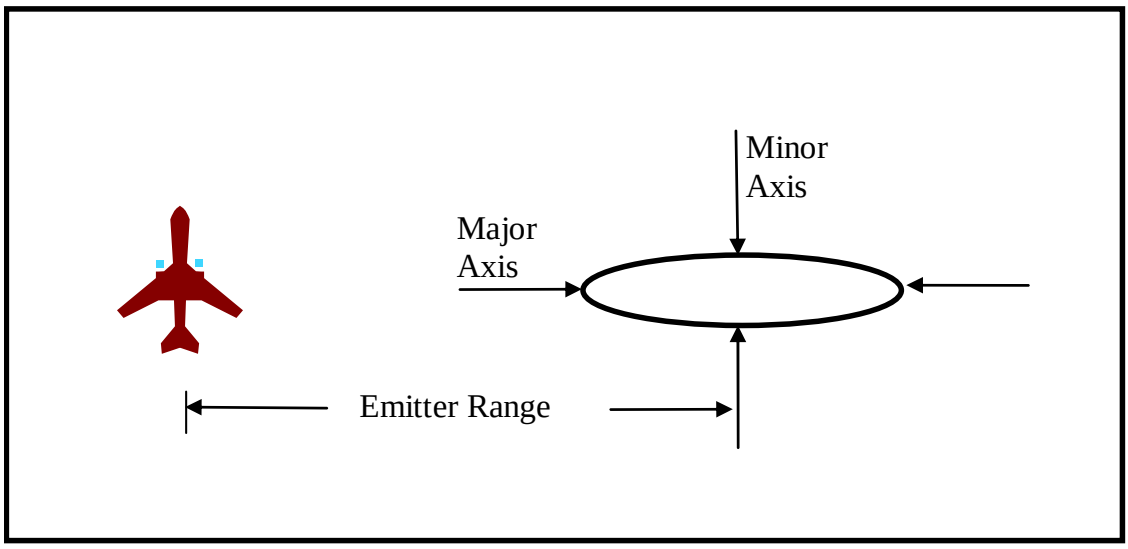


Figure B-4: Geolocation Error Ellipse.

Geolocation systems typically work in one of two ways. The first method is to track the calculated major and minor axes of the error ellipse until they collapse to a specified percentage of the estimated emitter range and when this occurs the system assigns the emitter a location at the centre of the error ellipse. The second method works the same way as the first method, but it does not stop computing the error ellipse and continues to update the computed position as long as the emitter is transmitting. The relevant difference from an analysis standpoint is that if the system determines a single location of an emitter, each location produces a discrete location error. While, if the system continuously computes emitter location the data will consist of a time-based series of emitter location errors.

One means of evaluating the performance of a system that continuously computes emitter location is a version of response time. The performance of the system is evaluated by determining the time for the major ellipse axis to collapse to a specified percentage of the range to the emitter; the time is a function of the geometry. Since there are many considerations consultation with a professional statistician is recommended.

B.5.2 Jammer MOPs

The ultimate measure of a jammer's utility is whether or not it can protect the aircraft it is designed to protect. This is exceedingly difficult to quantify, particularly in a flight test environment. Aircraft survivability presents a complex evaluation with many combinations and permutations, where jammer effectiveness is only one variable. Each engagement is unique and is a function of the specific conditions of the engagement. Other considerations include manoeuvres, tactics, and other countermeasures such as support jamming or chaff.

Some measures are relatively easy to quantify. Guided weapons or weapons direction systems must maintain an angular track on a target. Radar directed weapons also track targets in range and/or velocity. A means of evaluating the performance of a countermeasures system is to record the tracking error data associated with a target under non-jamming conditions, a condition known as dry and comparing them to the tracking error data collected under the same conditions with the countermeasures system operating, a condition known as wet. Another measure is to evaluate whether or not the jammer selected the correct technique.

While wet-to-dry track error comparisons are useful MOPs for analyzing EA technique effectiveness, they need to be used with caution, as different weapon systems have varying degrees of tolerance to track errors. Some systems can incur very large tracking errors and still successfully complete an engagement. Other MOPs that attempt to address more operationally relevant aspects of a jammer's performance are: cumulative missile miss distance comparisons, reduction in shot opportunities, and Reduction in Lethality (RiL). Each of these has strengths and weaknesses as well.

Simulated missile or projectile fly outs underlie a number of jammer MOPs. These simulations can be purely digitally modelled or use some combination of flight test generated radar data and modelled missile or projectile fly outs. EW data analysts need to fully understand the limitations of the models they use. One of the main EP features of modern radar systems is a well-trained operator in the loop. Understanding how the operator is represented in the model is vital to understanding its utility.

Historically, one of the major problems with using flight test data to support missile fly out modelling has been the inability to precisely and accurately know the location of the target aircraft. While OAR reference radars are good enough for many purposes, their accuracy imposed significant limitations on missile fly out simulations that attempted to determine hits or misses. The TSPI location errors for the test aircraft were often on the order of the warhead lethal radius, particularly for smaller missiles. This problem has been somewhat alleviated by the use of very accurate Global Positioning System (GPS) data as a Time-Space-Position Information (TSPI) source. Testers should remain aware of the importance of precise and accurate target TSPI data.

No single MOP comprehensively addresses the performance of an EA system; however, every good MOP indicates something about the performance of an EA system. A prudent analyst will examine as many MOPs as practical to evaluate the system performance.

B.5.2.1 Tracking Errors

Dry versus wet tracking errors are commonly presented in a range versus tracking error format; with the range separated into bins. Median errors are most commonly presented. Data are presented by threat system and test conditions and normally consist of a compilation of several individual passes. Figure B-5 shows an example of median range tracking error plot. Median is more commonly used than mean as an average since a small number of very large errors can cause misleadingly large errors if the mean is used.

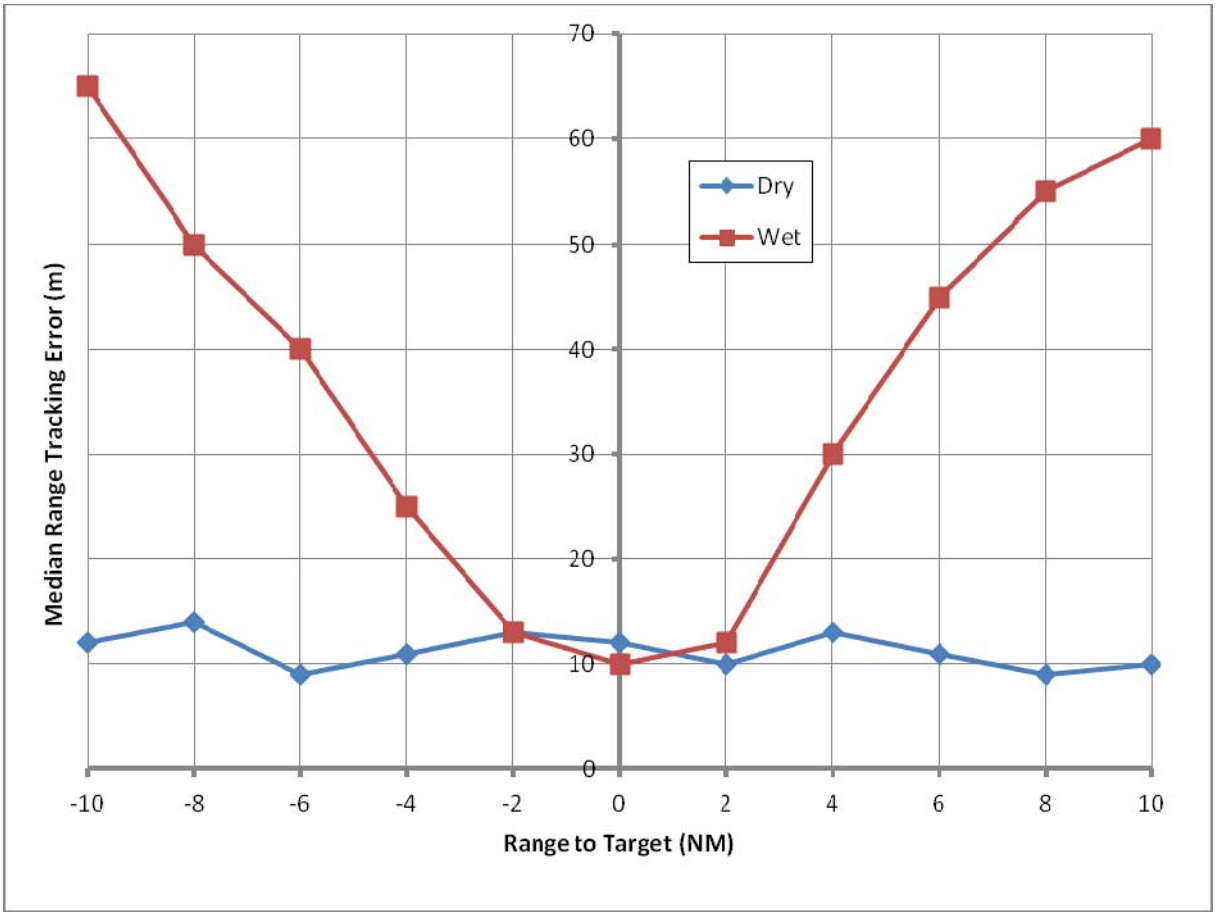


Figure B-5: Sample Median Range Tracking Error Plot (For a Given Threat Radar and Test Condition).

B.5.2.2 Cumulative Missile Miss Distances

Cumulative missile miss distance plots present the results of simulated missile fly outs as a comparison of dry versus wet results. Figure B-6 shows a sample graph. The graph indicates that jamming has increased the missile miss distance. Ninety percent of the dry run miss distances were within 10 meters while only 10 percent of the wet run miss distances were within 10 meters. The data should be collected to the maximum extent possible under the same conditions for both dry and wet runs.

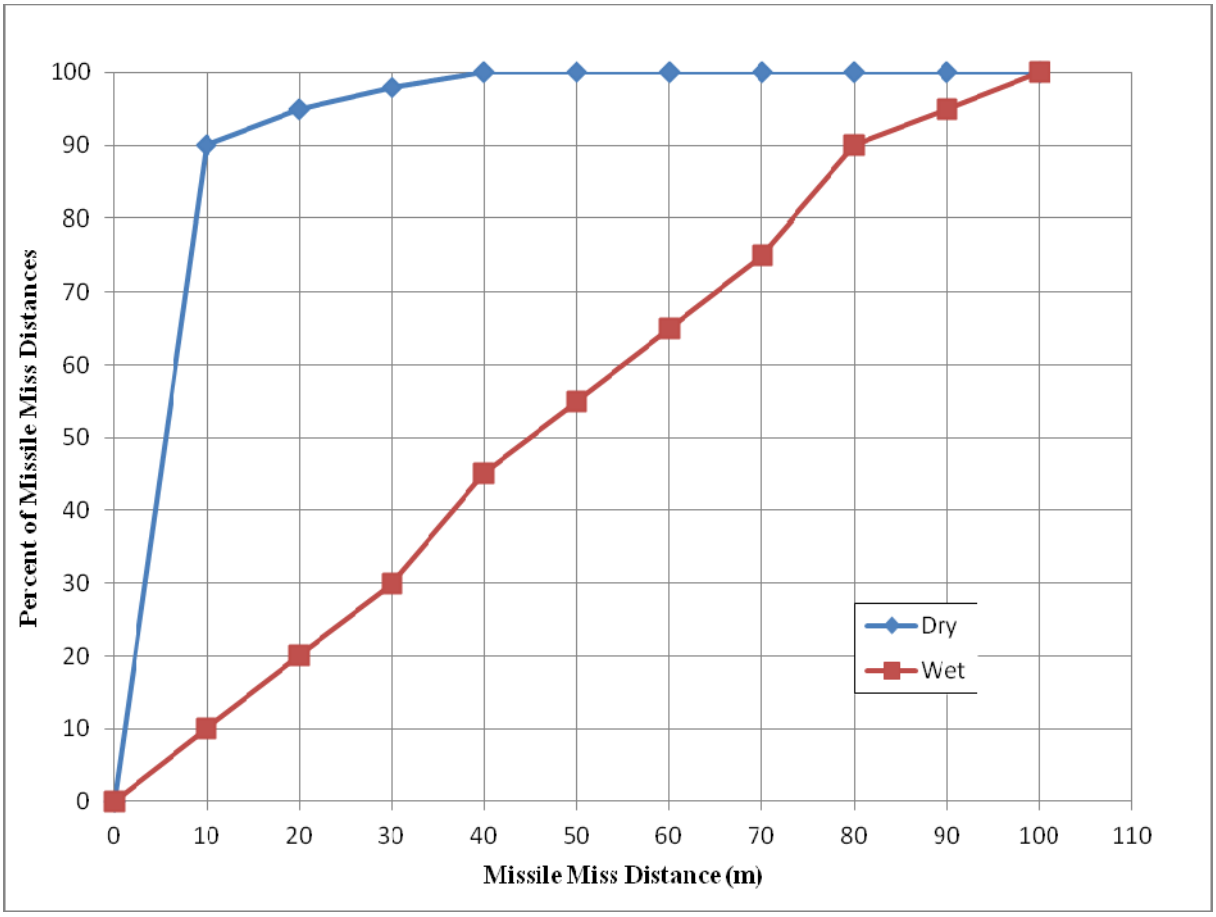


Figure B-6: Example of a Missile Miss Distance Cumulative Percentage Plot.

B.5.2.3 Reduction in Shot Opportunities

One of the benefits of effective self protection jamming is that the EA technique will disrupt the threat system and deny the threat system operators shot opportunities. Reduction in Shot Opportunities (RiS) can be expressed as:

$$RiS = \left[1 - \frac{\frac{Number\ of\ Shots\ (Wet)}{Number\ of\ Passes\ (Wet)}}{\frac{Number\ of\ Shots\ (Dry)}{Number\ of\ Passes\ (Dry)}}\right] \times 100$$

B.5.2.4 Reduction in Lethality

Reduction in Lethality (RiL) is a measure that attempts to quantify the effectiveness of the jammer. It is defined as follows:

$$RiL = \left[1 - \frac{\frac{Number\ of\ Hits\ (Wet)}{Number\ of\ Passes\ (Wet)}}{\frac{Number\ of\ Hits\ (Dry)}{Number\ of\ Passes\ (Dry)}}\right] \times 100$$

RiL has two main advantages: it is easy to compute and it focuses on whether or not the threat system successfully engaged the protected aircraft. However, it has a number of disadvantages. The primary

shortfall comes from determining the definition of a "hit". Hits are commonly determined by comparing the calculated or simulated missile miss distance to a predetermined miss distance from the aircraft. This distance is often based on the largest dimension of the aircraft (for example, half of the wing span) plus some fixed number representing the lethal radius of the warhead. This considerably oversimplifies the warhead-target interaction, particularly for missiles with small warheads. Another shortfall is that the term RiL is a misnomer; the expression defined above might more properly be termed a Reduction in Susceptibility, since it address hits and misses instead of kills or lethality. Additionally, when RiL is based on flight test data the previously discussed problem of target location accuracy and precision must be considered.

B.6 REFERENCES

[1] Defense Acquisition University, "Glossary of Defense Acquisition Acronyms and Terms", 12th edition, Defense Acquisition University Press, Fort Belvoir, Virginia 22060-5565, July 2005, p. B-40.

[2] DAU, p. B-103.

[3] DAU, p. B-103.

[4] DAU, p. B-103.

[5] Stevens, S.S., "On the Theory of Scales of Measurement", Science, Vol. 103, No. 2684, 7 June 1946, p. 677.

[6] Stevens, pp. 678-679.

B.7 FURTHER READING

Huber, A.F., Timmerman, T.J. and Gibbons, J., "Testing and Evaluating the Military Worth of Electronic Warfare Systems", Society of Flight Test Engineers, 28th Annual Symposium Proceedings, August 1997.

Sproles, N., "Avoiding Confusion – A Case for Establishing Standards for Measures of Effectiveness", ITEA Journal, September/October 1998, pp. 26-30.

Sproles, N., "Coming to Grips with Measures of Effectiveness", Systems Engineering, 1998, pp. 1-8.