Assessing Submarine Sonar Performance Using Statistically Designed Tests

George M. Khoury, Justace R. Clutter, and V. Bram Lillard

THE PROBLEM

Historical at-sea methods for determining Anti-Submarine Warfare performance of the Navy's submarine sonar system are unable to characterize performance across a range of operational conditions and yield statistically significant results.

To address the shortcomings of A-RCI at-sea testing, IDA proposed augmenting the at-sea operational test events with so-called Operator-In-the-Loop (OIL) laboratory tests.

The Acoustic Rapid Commercial-off-the-Shelf Insertion (A-RCI) is the Navy's newest submarine sonar processing system. It provides hardware and software to process data from the submarine's sonar arrays and display those data to the sonar operators. A-RCI uses a spiral development model to procure new, commercial off-the-shelf computing hardware every two years. Buying new computing hardware over time capitalizes on the decreasing cost of processing power and ensures that an acceptable balance between obsolescent and modern hardware is maintained. To take advantage of the ever-improving processing power from hardware upgrades, a new version of A-RCI software, denoted an Advanced Processing

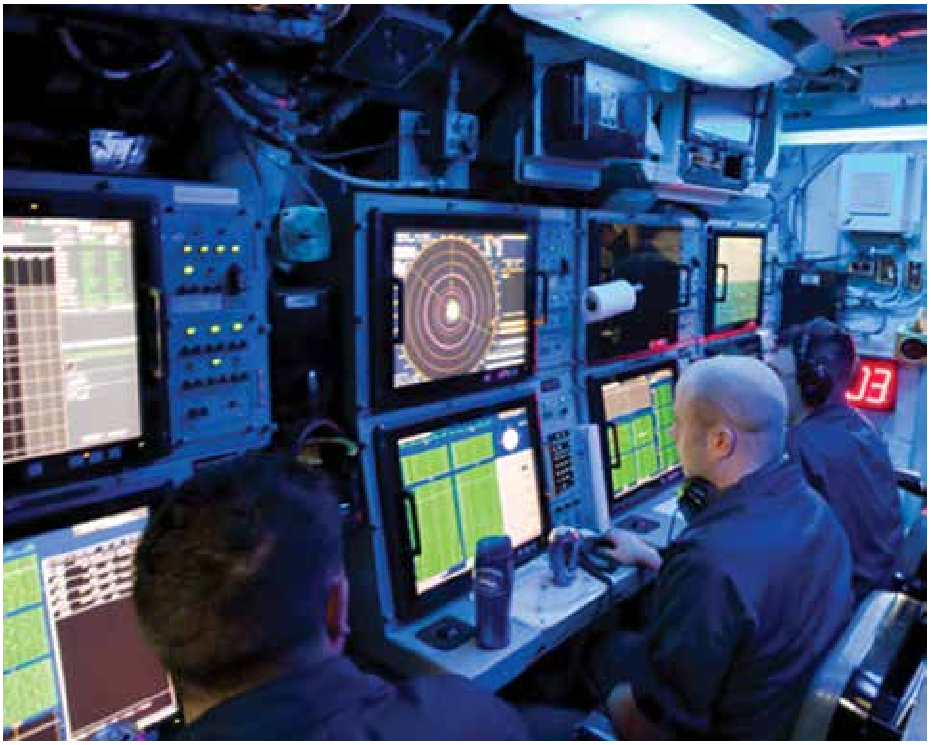


Photo by Bryan Jones https://www.flickr.com/photos/bwjones/3552816442

Figure 1. Four A-RCI Sonar Consoles aboard a Submarine

Build (APB), is developed every other year. Each APB incorporates feedback from Fleet users, fixes bugs discovered in previous versions, and adds new algorithms developed by industry and academia.

The primary role for A-RCI is to manage the large amount of information coming from the sonar arrays and display it to the operator so that he can make sense of it. To understand the scale of the operator's problem, consider that a *Virginia*-class submarine uses six sonar arrays for submarine searches, each providing information on all bearings, multiple elevation angles, and a range of frequencies. The sonar operator must monitor this multi-dimensional search space constantly, and it is impossible to display all of the information simultaneously. A-RCI provides displays and automation to help the operators manage this search space and help them detect contacts as quickly as possible.

The Navy's primary metric with which to evaluate A-RCI performance in the Anti-Submarine Warfare (ASW) mission is denoted ΔT. It is defined as the median time it takes for an operator to detect a submarine contact once that submarine's signal becomes available for display on sonar system screens. Although ΔT is not a measure of the submarine's overall ASW capability, it does quantify A-RCI's role in the detection process. The ongoing goal of A-RCI processing improvements is to minimize the time needed to find target signatures.

At-sea tests of A-RCI consist of two submarines searching for each other in a specified area. Although this technique provides an operationally realistic environment, it suffers from several drawbacks. Most notably, at-sea testing has never been able to show a statistically significant improvement in A-RCI performance over the course of a decade, during which time many software and hardware upgrades were fielded to the Fleet. A comparison has been impossible because two software versions are never compared in the same at-sea event, and the results of a test can depend on target and environmental characteristics that are impossible to control. Additionally, at-sea testing uses a single target and a single operational environment, which limits the assessment of performance of the new APB to only a small portion of the operational envelope. Finally, the cost and variability of at-sea testing have resulted in poor quantification of APB performance in the conditions tested.

To address the shortcomings of A-RCI at-sea testing, IDA proposed augmenting the at-sea operational test events with so-called Operator-In-the-Loop (OIL) laboratory tests. In an OIL test, a Fleet operator sits at a laboratory mockup of the A-RCI sonar system. The laboratory then plays back a recorded at-sea encounter between two submarines, and the operator declares when he has detected the threat submarine.[1] The laboratory allows the same encounter to be replayed on different

[1] U.S. submarines are capable of recording raw sonar data, that is, the voltage recorded by the individual hydrophones that make up the sonar arrays. Because these raw data are recorded before they are processed by A-RCI, it is possible to process the recorded data on any version of A-RCI.

versions of A-RCI, which perfectly controls for environmental and target variability; the only difference between the two presentations is the software used to process the data. The primary limitation of the OIL testing is that it only allows for a single array to be processed at one time. Therefore, the sonar array to be processed needs to be a controlled test factor, whereas in real operations all arrays operate simultaneously.

For several years, the Navy has used a similar laboratory test method to compare new versions of A-RCI to old versions, but typically used only a few submarine encounters for each comparison, and published the results long after the software was fielded. As part of our support to DOT&E, IDA proposed expanding the scope of such tests to include a wider variety of test conditions, as shown in Table 1, and to employ Design of Experiments methodologies to generate a more robust test that would characterize performance across a range of operational conditions. The primary goal of the test was to compare the latest version of the sonar system, denoted APB-11, with the previous version, APB-09. To characterize the systems, the test used operators of varying proficiency and controlled for characteristics of the target and the array being used.

FACTOR LEVELS HYPOTHESIZED EFFECT

IDA developed a 120-run, D-optimal, split-plot test design, with the distribution of runs as shown in Figure 2. A "run" consists of a single operator viewing a single recorded encounter, and a "Null" run is one in which no target is present. The

Table 1. Factors and Levels Used in the OIL Testing Analysis

Factor	Levels	Hypothesized Effect
Target Type	SSN, SSK	SSNs and SSKs exhibit different acoustic signatures. SSNs typically have more discrete tonal information because of the machinery associated with the nuclear reactor.
Array Type	A, B	Array type A typically detects targets at longer ranges, which would be expected to generate larger ΔTs.
Target Noise	Loud, Quiet	Loud targets are detected at longer ranges, which could lead to longer ΔTs. Conversely, loud targets typically have more discrete tonal information and are easier to identify, which could result in shorter ΔTs.
APB Version	APB-09, APB-11	The primary goal of the test was to compare the latest version of the sonar system, APB-11, with the previous version, APB-09.
Operator Proficiency	1 to 20	More proficient operators will detect a submarine more quickly. The numeric scale was developed by the Naval Undersea Warfare Center and is based on an operator's experience with the A-RCI system.

	Target Type		Array B	Array A
APB-11	SSK	Quiet	6	6
		Loud	6	12
	SSN	Quiet	6	12
		Loud	6	12
		Null	6	

	Target Type		Array B	Array A
APB-09	SSK	Quiet	4	4
		Loud	4	8
	SSN	Quiet	4	8
		Loud	4	8
		Null	4	

Total Number of Runs

72

48

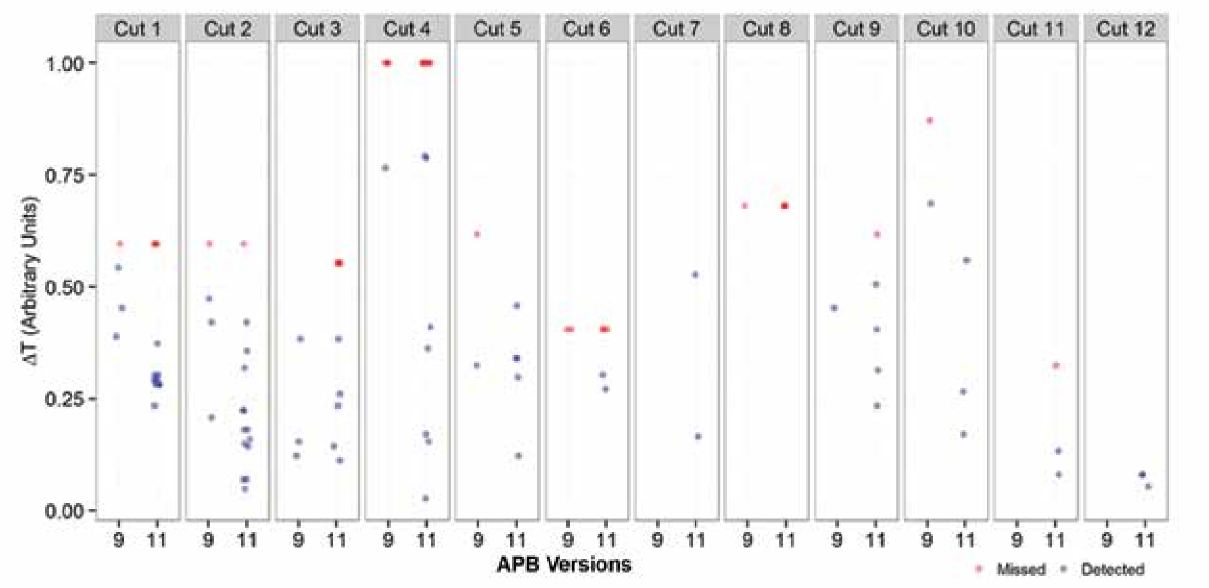
120

Figure 2. OIL Test Design Matrix

split-plot structure was used to limit the number of changes between the APB versions, as each change of APB required approximately 12 hours. A considerable amount of replication was built into the design to account for the fact that operator proficiency was not explicitly controlled. Instead, operators were chosen at random, and their proficiency was recorded during the events, which ensured a balanced distribution of proficiencies. Each operator reviewed up to six tapes, including a blank tape to check for false alarm rate. Finally, the Navy desired to focus the testing on APB-11, which resulted in the asymmetric test design shown; while this was not optimal for determining whether a significant APB difference existed, it provided a more precise understanding of performance for APB-11 (tighter confidence intervals).

TEST RESULTS

Figure 3 shows the raw results of the test. Each panel shows the results for a recorded encounter, with APB-09 results on the left and APB-11 results on the right. The blue dots are detection times; the red dots indicate runs in which the operator never detected the target before the



Each panel (Cut 1, Cut 2, ...) shows the results for a single recording. Blue points indicate detection times (arbitrary units). Red points indicate runs in which the operator did not detect the target; the time in these cases is the length of the recording.

Figure 3. Raw Results from A-RCI OIL Testing

recording finished. The location of the red dot indicates how long the target was on tape and not detected.

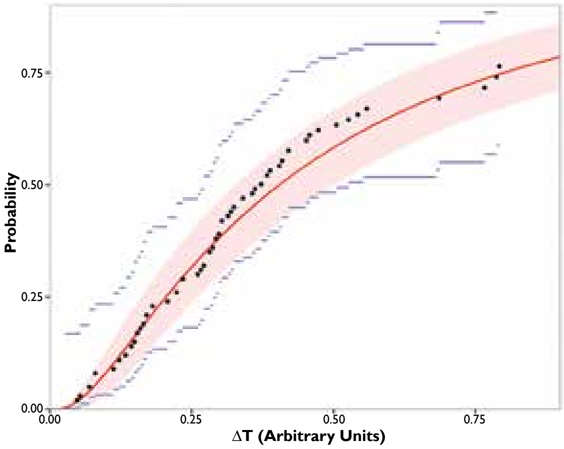
The advantage of examining the results by recording is that recordings control all aspects of the encounter; the environment and target are exactly the same for each playback, so any difference in performance is due to either operator proficiency or the capability of the processing system. Since the test was well balanced in terms of operator proficiency, any observed differences are most likely due to the processing system. In general, APB-11 exhibited improved performance in almost all of the recorded encounters; in each panel, the dots are generally lower for APB-11 than they are for APB-09, reflecting shorter times to detect threat submarines. Therefore, even without statistical analysis, APB-11 appears to be an improvement over APB-09. Such a limited analysis does not, however, make use of all the available information; APB-11 appears to be better, but the improvement varies with recording and it is unclear why. The test was designed to determine which of the controlled factors affect A-RCI performance, and for that a statistical analysis is necessary.

We performed a regression analysis to better understand how the controlled factors affected A-RCI performance. Our analysis accounts for missed detections by treating them as censored data points; in these cases, we assumed that the operator would have detected the contact if given enough time, so the full recorded length of time the contact was on the display is used as a lower bound estimate for the ΔT. We assumed that the data followed a lognormal distribution, in which the probability of observing a detection time x is the following:

$$p(x\,|\,\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

Here, μ is related to the median of the distribution, and σ is a measure of its spread. Making this assumption allowed us to incorporate the missed detections using standard censored data analysis techniques.

Although there is no *a priori* reason why the data should follow a lognormal distribution, our initial assumption was well supported by the data. Figure 4 shows the empirical cumulative distribution function of the data, along with a lognormal fit in red, the confidence region on the lognormal fit in pink,



Red line shows a lognormal fit. Pink region shows the 80% confidence region on the lognormal fit. Blue lines indicate the 80% confidence region on a non-parametric fit to the distribution. The data are well described by a lognormal distribution.

Figure 4. Empirical Cumulative Distribution of the OIL Data

and the confidence region of a non-parametric fit in blue lines. The data appear to be well described by a lognormal distribution.

Next, we assigned each recording to the factors listed in Table 1, and then fit the data according to the following model:

$$x \sim lognormal(\mu,\sigma)$$

$$\sigma = \text{constant}$$

$$\mu = \beta_0 + \beta_1\ OP + \beta_2\ APB + \beta_3\ Target + \beta_4\ Noise + \beta_5\ Array + \beta_6\ Target * Noise + \beta_7\ Target * Array + \beta_8\ Noise * Array + \beta_9\ Target * Noise * Array$$

That is, we assumed that the median detection time depends on the factors listed in Table 1, along with second and third order interactions, and that the σ parameter was constant. In fact, we examined many possible models, including those with variable σ, but this model resulted in the lowest Akaike's Information Criterion (AIC), a metric of model desirability. Table 2 shows the results of the final fit and describes the qualitative behavior of the coefficients. All of the first-order effects were highly significant. Notably, APB-11 performed significantly better than APB-09, holding all other effects equal – and the magnitude of the effect was

Table 2. Results of the Model Fit to the Data

Term	Value†	Description of the Effect
β_1 (Operator experience level)	-0.074 ± 0.041	Increased operator proficiency results in shorter detection times. An increase in proficiency of one unit reduces median detection time by 7 percent.
β_2 (APB)	0.307 ± 0.129	Detection time is shorter for APB-11, **by 46 percent.**#
β_3 (Target)	0.359 ± 0.126	Detection time is shorter for SSN targets.
β_4 (Noise)	-0.324 ± 0.125	Detection time is shorter for loud targets.
β_5 (Array)	0.347 ± 0.125	Detection time is shorter for the Type B array.
β_6 (Target*Noise)	0.186 ± 0.126	Additional model terms added to improve predictions. The third-order interaction is marginally significant, so all second order interactions nested within the third order interaction were retained to preserve model hierarchy.
β_7(Target*Array)	0.011 ± 0.125	
β_8(Noise*Array)	0.021 ± 0.126	
β_9 (Target*Noise*Array)	-0.180 ± 0.125	

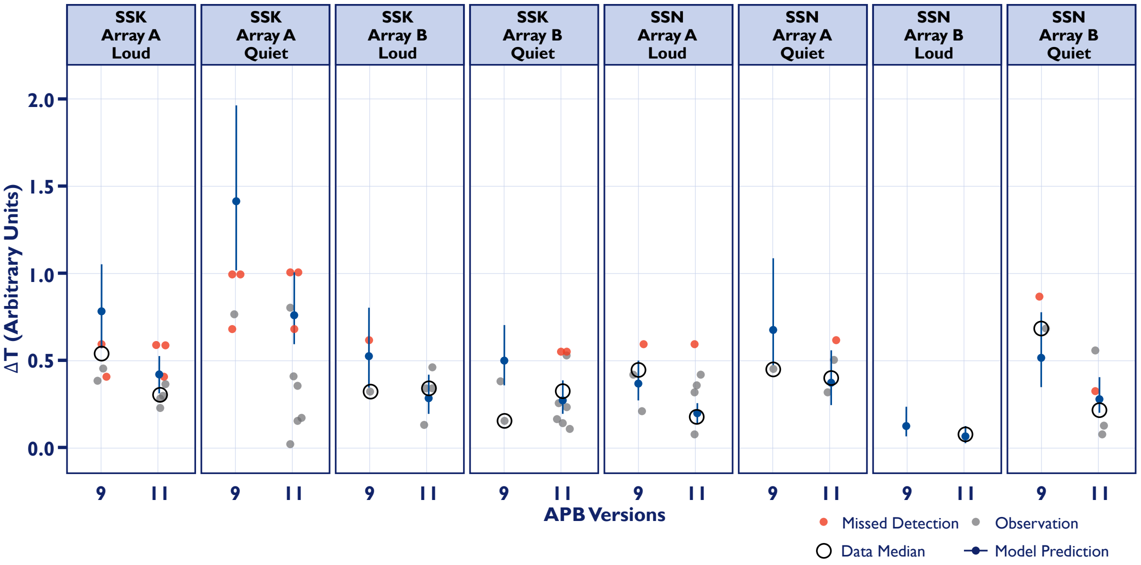
† Confidence interval is an 80% Wald interval

# APB-11 Provides a Statistically Significant Improvement.

substantial, as APB-11 detection times were 46 percent shorter than APB-11 times. Also, APB had no interaction with the other factors, which means APB-11 produced the same improvement regardless of the other factors. It did not matter whether the target was loud or quiet, SSN or SSK; switching from APB-09 to APB-11 reduced the median detection time by approximately 46 percent. This was the first time operational testing of A-RCI had shown a statistically significant improvement in an APB.

Figure 5 shows the results of the model fit (blue dots, with 80% confidence intervals shown as vertical lines), along with the actual median detection times in each group (black) and the raw detection times (light blue and red, as before). The model predictions generally agree with the median in each bin, indicating that our relatively simple model provides a good fit to the data. There is, however, notable disagreement between the data median and the model prediction for one bin: quiet SSK targets with array type B in APB-09. The difference is due to sparse data, rather than a poorly fitting model. The data median in this case is based on only three data points and is therefore highly variable, making it a poor estimator of the true performance in that bin. We believe the model estimate predicts the performance that would be observed if additional runs were conducted with APB-09.

Our analysis provides several benefits over the less sophisticated analysis based solely on individual recordings. First, differences in performance are now attributable to operationally relevant factors, such as target type or array type. In contrast to the naïve analysis by recording, our statistical analysis shows that APB-11



The model fits the data well and indicates that APB-11 outperforms APB-09 in all conditions. Data medians were omitted when the data in the bin were inadequate to support the estimation of a median value (e.g., too few data points). This illustrates another advantage of using the empirical statistical model, since it can estimate median performance in every bin whereas traditional data analysis methods might not be able to provide a robust estimate.

Figure 5. Model Predictions (Blue), along with the Median Detection Time Observed in Each Bin

outperforms APB-09 by 46 percent on average across all conditions. Second, our analysis allows us to extrapolate to areas where the data are limited. A few of the experimental configurations presented in Figure 5 do not have an observed data median for comparison with the model prediction, either because there were few data points or because there was an excess of censored values. An analysis using a simpler technique would not have been able to estimate performance in regions where the data were inadequate to produce an estimate of performance.

CONCLUSIONS

Operator-in-the-Loop testing has proven to be an effective way to compare the performance of different versions of the sonar processing system and to discover how performance varies across a variety of operationally important factors. By playing back recorded data from real-world submarine encounters, OIL testing controls for target and environmental variability in a way that traditional at-sea testing cannot. It provides more data at a lower cost, which has enabled IDA to show a statistically significant improvement in A-RCI for the first time, and it has allowed us to quantify the operational factors that affect the improvement. Laboratory testing will not soon replace all at-sea testing, but it is a valuable complement.

Dr. Khoury is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of California, Santa Barbara.

Dr. Clutter is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of Kansas.

Dr. Lillard is an Assistant Director in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of Maryland.