

---

# Test Science Motivation and Report Guidance for DOT&E Action Officers

Dr. Catherine Warner


Science Advisor to the Director,  
Operational Test and Evaluation (DOT&E)





# DOT&E Guidance

## Dr. Gilmore's October 19, 2010 Memo to OTAs

 OFFICE OF THE SECRETARY OF DEFENSE  
1700 DEFENSE PENTAGON  
WASHINGTON, DC 20301-1700

OCT 19 2010


OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND  
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE  
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER  
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY  
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND  
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND  
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE  
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE  
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY  
DOT&E STAFF

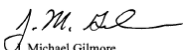
SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.



reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

  
J. Michael Gilmore  
Director

cc:  
DDT&E

2

- The goal of the experiment.** This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.
- Statistical measures of merit (power and confidence)** on the relevant response variables for which it makes sense. These statistical measures are important to understanding "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.



# Flawed Application of DOE to OT&E

1. **Clear test goals**
  - Focus on characterization of performance, vice testing to specific requirements
2. **Mission oriented metrics**
  - Not rigidly adhering to requirements documents
  - Continuous metrics
3. **Do not limit factors to those in requirements documents**
4. **Strategically control factors**
5. **Avoid confounding factors**
6. **Avoid single hypothesis tests**
7. **& 8. Consider all factors**
  - Understand that adding/removing factors does not necessary increase/decrease the size of the test.



OFFICE OF THE SECRETARY OF DEFENSE  
1700 DEFENSE PENTAGON  
WASHINGTON, DC 20301-1700

JUN 26 2013

MEMORANDUM FOR COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE (COMOPTEVFOR)

SUBJECT: Flawed Application of Design of Experiments (DOE) to Operational Test and Evaluation (OT&E)

In October 2010 I communicated my expectations regarding the use of DOE for developing rigorous, adequate, and defensible test programs and for evaluating their results. Over the past several years, all of the operational test agencies have implemented DOE practices to varying degrees and have offered training to their staff on the statistical principles of DOE. However, I am concerned that OPTEVFOR is not complying with the intent of the use of DOE as a method for test planning, execution, and evaluation. I find that most test designs focus exclusively on verifying threshold requirements, rely too heavily on hypothesis tests for test sizing, and all too often do not embrace the statistical tenets of DOE. Furthermore, OPTEVFOR has not updated its data analysis practices to capitalize on the benefits of using DOE.

One of the most important goals of operational testing is to characterize a system's (or system of systems') end-to-end mission effectiveness over the operational envelope. Such characterization of performance informs the Fleet and the system operators of its capabilities and limitations in the various conditions that will be encountered during combat operations. The goal of operational testing is not solely to verify that a threshold requirement has been met in a single or static set of conditions. I advocate the use of experimental design (DOE) to ensure that test programs (including integrated testing where appropriate) are able to determine the effect of *factors* on a comprehensive set of *operational mission-focused* and *quantitative* response variables. The determination of whether requirements have been met is also a test goal, but should be viewed as a subset of this larger and much more important goal.

Test designs and integrated evaluation frameworks (IEFs) developed by your staff will improve by following the direction provided in the remainder of this memorandum.

**I. A clear test goal must be created for each phase of test.**

As I state in previous guidance, as well as in the recently promulgated Test and Evaluation Master Plan (TEMP) Guide, a successful test plan must identify the goal of the test. Goals should be clearly identified in the TEMP as well as the test plan, and should be specific. Future test plans must state clearly that data are being collected to measure a particular response variable (possibly more than one), in order to characterize the system's performance by examining the effects of multiple factors. Test plans must also clearly delineate what statistical model (e.g., main effects and interactions) is motivating the strategic factor variation of the test.





# Assessing Statistical Adequacy of Experimental Designs in OT&E



OFFICE OF THE SECRETARY OF DEFENSE  
1700 DEFENSE PENTAGON  
WASHINGTON, DC 20301-1700

JUL 23 2015

OPERATIONAL TEST  
AND EVALUATION

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION  
COMMAND  
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND  
EVALUATION ACTIVITY  
COMMANDER, OPERATIONAL TEST AND EVALUATION  
FORCE  
COMMANDER, AIR FORCE OPERATIONAL TEST AND  
EVALUATION CENTER  
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used  
in Operational Test and Evaluation

Recent discussions within the test community have revealed that there are some misunderstandings of what DOT&E advocates regarding the appropriate use of statistical power when designing operational tests. I, as well as others in the test community, have observed that power calculations based on a single-hypothesis test on the overall mean are being used inappropriately by both government and industry in attempt to right-size a test. The purpose of this memorandum is to make clear what I view are best practices for the use of power calculations, as well as other statistical measures of merit that should be used to determine the adequacy of a test design.

Single-hypothesis test power calculations are generally inappropriate for right-sizing operational tests because they are not consistent with the goal of operational testing: to characterize a system's performance across the operational envelope. Furthermore, such estimates of power are unable to distinguish between both good and flawed test designs because they focus solely on the number of test points and ignore the placement of those points in the operational envelope. More informative power estimates exist. Power calculations that estimate the ability of the test to detect differences in performance amongst the conditions of the test (factors) will distinguish between good and flawed designs.

These "factor-level" power calculations are inherently related to the goal of the test; they not only describe the risk in concluding a factor is not important when it really is, but they are also directly related to the precision we will have on the quantitative estimates of system performance. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power to detect factor effects might fail to detect true system flaws; if it does, we have failed in our duty as testers.



- Re-emphasizes the importance of statistical power when used correctly.
- Highlights the importance of:
  - Clearly identifying a test goal
  - Linking the design strategy to the test goal
  - Assessing the adequacy of the design in the context of the overarching goal
- Highlights other quantitative measures of statistical test adequacy
  - Correlation
  - Variance of Predictions



# General Guidance

---

- **Tests conducted using Design of Experiments methodologies are now the standard**
- **DOT&E Reports are including more content on analysis methodologies**
- **Report guidance**
  - Focus on operational impact!
  - Use graphs to illustrate key findings
  - Use only enough statistical jargon to provide a broad overview of the analysis for interested audiences



# Suggested Content

---

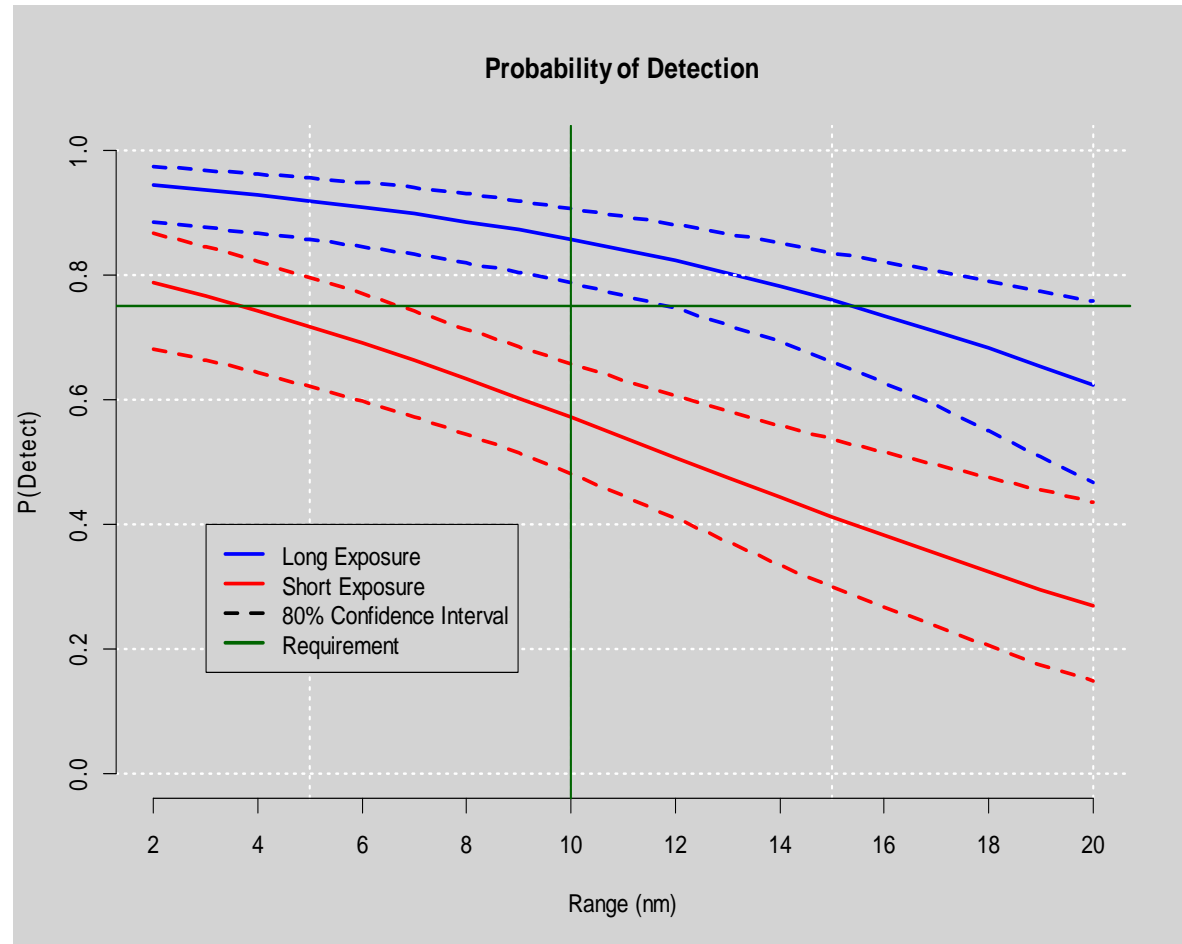
- **Executive Summary and Main Report Body**
  - High level conclusions
  - A widely understandable explanation of the results, including significant/notable factors or interactions
    - » Focus is on what the data are telling us about performance and the system's mission effectiveness
    - » Focus is not the statistics themselves
  - One or more summary graphs or charts that clearly depict the most important results
- **Footnotes in the main body:**
  - Brief explanation of the statistical test or method used to obtain the results
  - p-value(s) from the statistical test used
  - Other basic definitions or explanations that would help the more technically-oriented reader understand how and why certain conclusions were drawn
- **Appendix (if needed):**
  - Explanation and discussion of a statistical technique that is more complex or involved than basic statistical modeling (e.g. Bayesian analysis, mixed models, etc.)
  - Discussion of the statistical model selection process (i.e. if a large number of factors were considered, how the final model terms were decided)
  - Charts and graphs depicting modeling results, if too numerous for the main body
  - Residual analysis / other model validation results, graphs, and discussion



# Automatic Radar Periscope Detection and Discrimination (ARPDD) Notional Example

- **Graph is an example of the analysis that was included in the DOT&E report.**
- **Key conclusions:**
  - Periscope exposure time was a primary factor in ARPDD's ability to detect
  - Only at very short range for the shorter exposure time was the system able to meet the requirement (two-factor interaction)

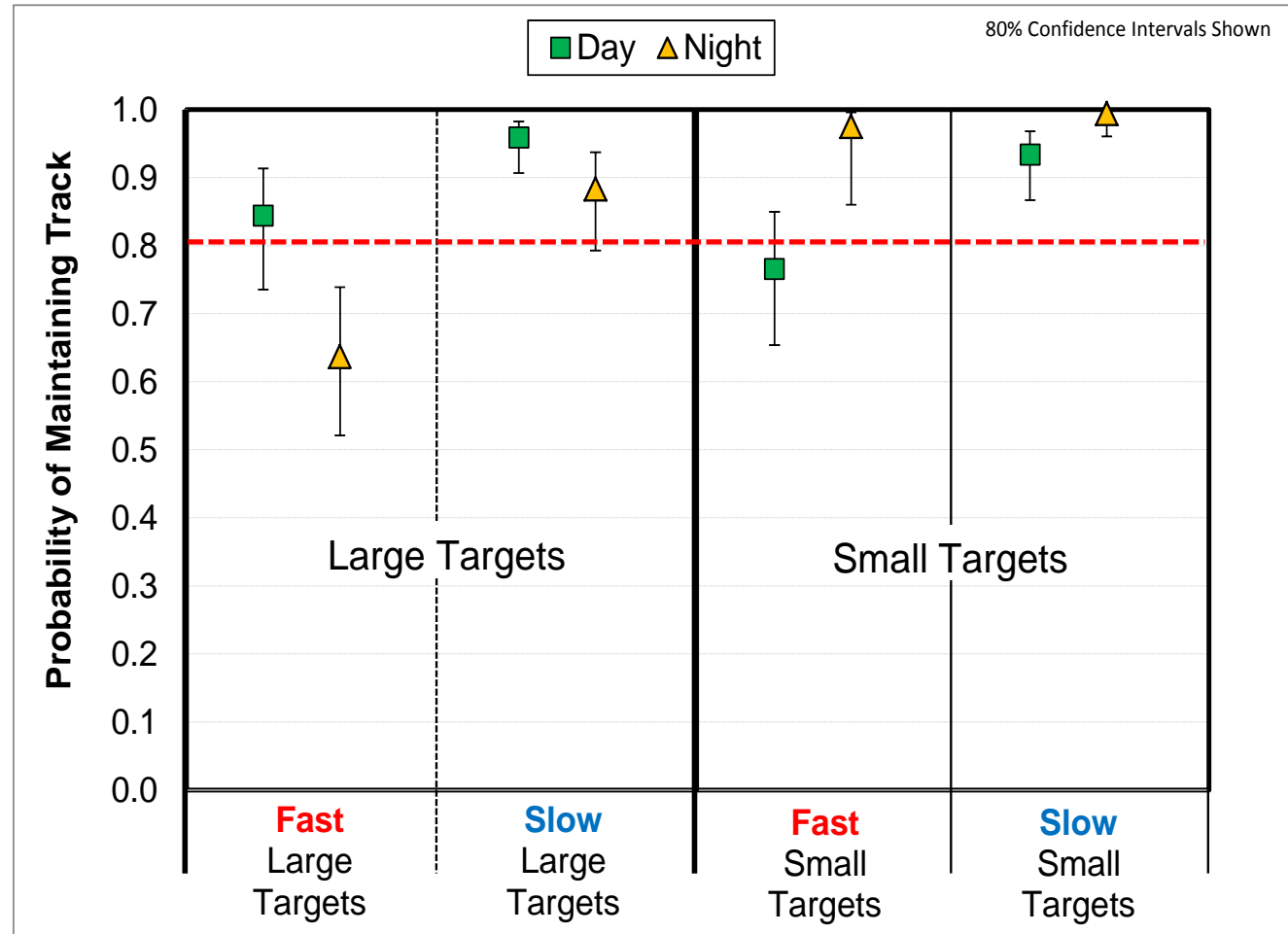
Graph is for illustration only, the actual numbers are not provided





# Analysis of Optical Tracking System (probability of maintaining track)

- Analysis revealed areas of degraded performance that would have otherwise been missed.
- Analysis enables performance characterization across multiple conditions







# AH-64E FOT&E I

- DOE executed close to plan

	Battlefield Density	Low		High	
		Day	Night	Day	Night
	Light	Day	Night	Day	Night
L16 Targeting Data	no	3	1	2	2
	yes	6	2	3	3

*Cells indicate missions executed per condition*

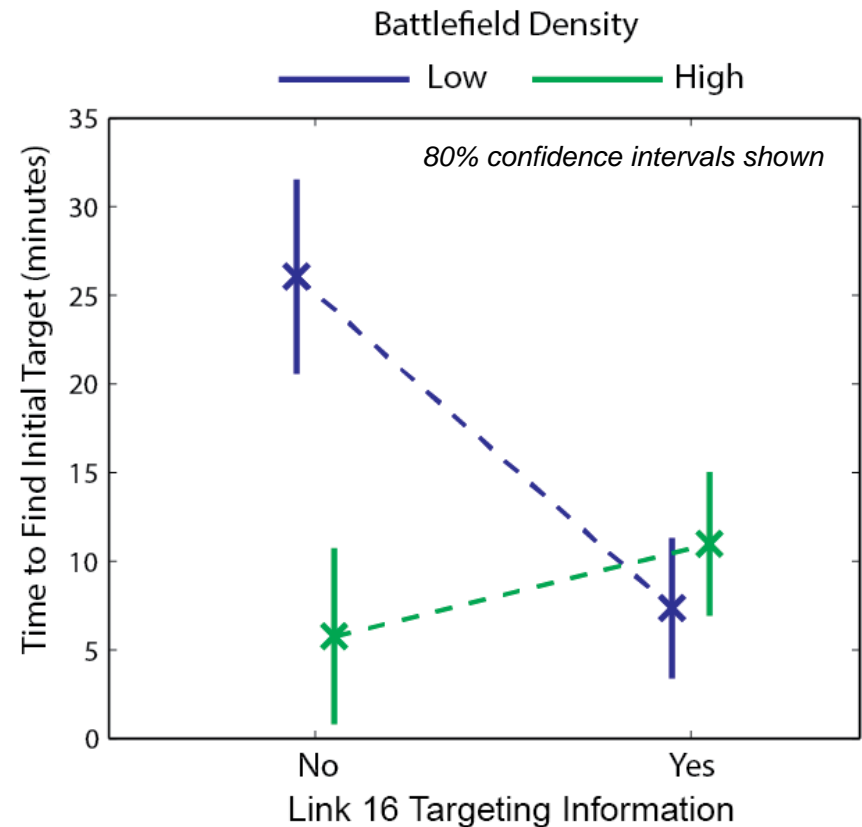
- **Statistical Result**

- L16 targeting data, battlefield density were statistically significant; light was not.
- Two factor interaction between BF density and L16 targeting data was significant

- **Bottom Line Result**

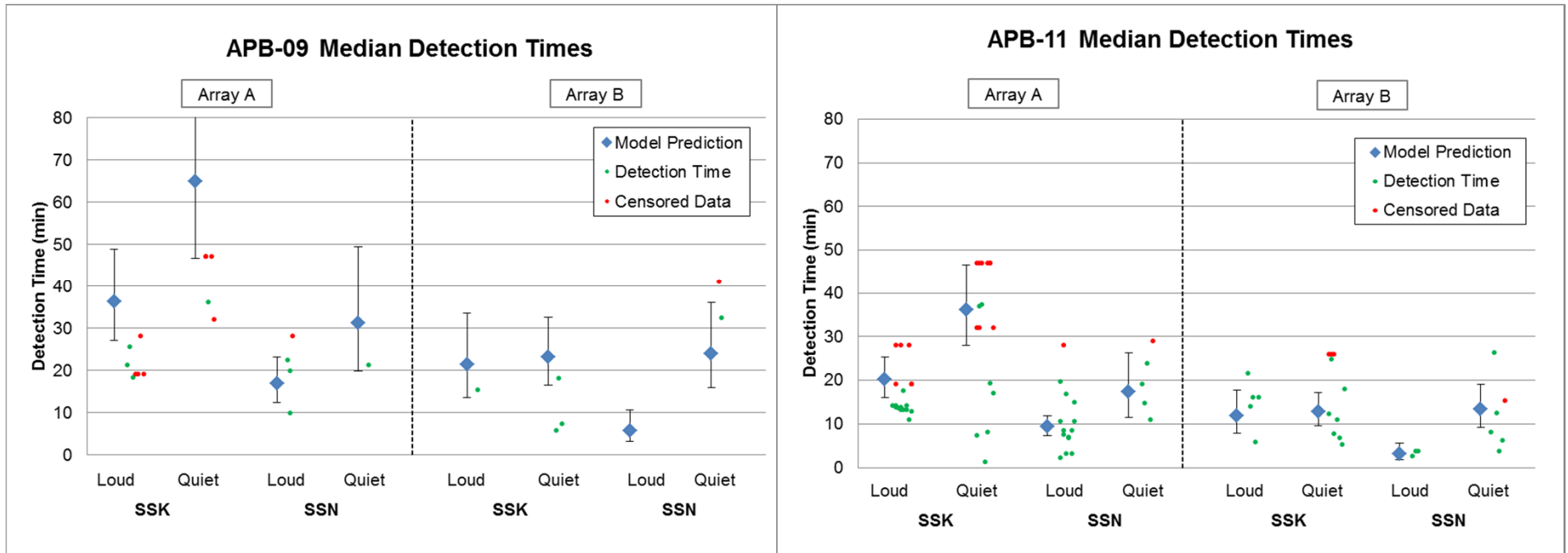
- L16 has a bigger effect on low density battlefields
- It is easy to find a target on a high density battlefield

- Graph shows interaction between factors





# Submarine Detection Time



- **Compares detection time between two different software versions**
  - Median detection times show a clear advantage of APB-11 over the legacy APB
- **Performance differences across different operational conditions are statistically significant**



## Q-53

- **Probability of detection for the Q-53 counterfire radar using the 360 degree operating mode against single-fired artillery projectiles is highly depended on the shot trajectory**
  - Interaction effect shown below indicates for high trajectories probability of detection is relatively constant
  - For low trajectories there is a large reduction in detection for weapons that are further away from the radar.

