SYM-AM-18-043

PROCEEDINGS
OF THE
FIFTEENTH ANNUAL ACQUISITION RESEARCH SYMPOSIUM

WEDNESDAY SESSIONS
VOLUME I

Acquisition Research:
Creating Synergy for Informed Change

May 9–10, 2018

Published April 30, 2018

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

Using Developmental T&E to Inform Operational T&E Decision-Based Analysis

Dashi Singham—is a Research Associate Professor of Operations Research at the Naval Postgraduate School, where she researches, teaches, and advises student theses. Dr. Singham's primary areas of focus include simulation modeling, simulation analysis, and applied statistics, with most of her work on developing new methods and metrics for analyzing simulation output. Her areas of application include energy and intelligence systems. She received her PhD in Industrial Engineering and Operations Research from the University of California Berkeley in 2010. [dsingham@nps.edu]

Abstract

We describe two-stage sequential experiments that are used in building and testing valid simulation models. In the first stage, preliminary samples are taken to estimate performance and inform the parameters for the experiments in the second stage. These two-stage experiments can be mapped to test and evaluation (T&E) by having the first stage applied to developmental test and evaluation (DT&E) and the second stage applied to operational test and evaluation (OT&E). By considering DT&E and OT&E as part of a combined two-stage experiment, we can better leverage the results of DT&E to inform OT&E.

Introduction

Statistical experimentation in test and evaluation is critical to obtaining clear, valid results and recommendations regarding the quality of a system being tested. We refer to test and evaluation (T&E) of a system, where a system can be a weapon, computer program, piece of machinery, and so forth. While much of the methodology for T&E has been developed, there is still much room for improvement in terms of ensuring widespread knowledge and implementation of statistical methods. Hill (2017) states, "The current T&E workforce, while very competent in the engineering domain and mechanics of test, will benefit by improving their baseline level of statistics, their statistical fluency, thus firming up their overall knowledge base" (p. 123).

This research develops two-stage statistical procedures that use developmental test and evaluation (DT&E) data to design and conduct operational test and evaluation (OT&E) plans. Two-stage procedures rely on data collected in a first stage to estimate key parameters that are needed to determine what types of future tests should be run to answer a research question. In a T&E setting, these estimated parameters have some uncertainty given that testing conditions may be limited in the first stage or approximated using simulation. This uncertainty can be used to determine what tests and statistical parameters to use in the second stage. For example, if DT&E reveals strong performance in some areas and weaker performance in others, we can design OT&E tests that allocate more effort to quantifying the effect of the weaker performance areas on overall system sustainability.

Two-stage statistical procedures are commonly used in analyzing simulation models. The first stage runs some preliminary experiments to estimate key parameters, like the variance and distribution of the output. Then, second-stage experimental parameters are chosen and the results from the second experiment contribute to the final assessment of the system. This research draws on two-stage procedures by mapping first-stage methods to DT&E, where simulation or less-costly experimental methods are available. The second-stage method is then mapped to OT&E with an emphasis on the fact that these experiments may be much more costly.

Examples of highly cited two-stage procedures include Chick and Inoue (2001) and those reviewed in Goldsman and Nelson (1998). These methods use the first-stage samples to estimate the variance, among other parameters, of multiple systems. Estimation of system variance is critical to determining the details of an OT&E experiment. Giadrosich (1995) describes how an estimate of the standard deviation can be used to choose the sample size, and sequential sampling methods that rely on this variance estimation are presented in Singham (2014). We note that much of the simulation literature now focuses on fully sequential sampling rather than two-stage sampling, but these fully sequential methods may not always be appropriate for T&E because of high sampling costs and potential for bias.

This paper exploits two-stage statistical procedures to provide a better link between statistical methods used in DT&E and OT&E testing. The case study presented addresses the unique challenges present within a T&E environment, such as specific capabilities requirements, limited budgets, and risk associated with an incorrect evaluation. OT&E often requires a much higher budget due to the operational nature of the testing. Thus, the information from the first-stage is critical in determining where effort should be focused in the second stage. However, in some cases, sophisticated simulation models can be employed for integrating testing, combining aspects of developmental and operational testing. For example, Allen (2010) describes the Boeing Engineering Development Simulator in its ability to replicate many operational settings while testing the enhanced capabilities of the aircraft, saving costs by using a simulated environment.

The next sections summarize the background in T&E and two-stage procedures, present a proposed two-stage algorithm, and apply the algorithm to a case study.

Background

DT&E and OT&E each pose their own set of unique challenges. DT&E is often performed under highly controlled or even simulated environments, so there are limitations on how much this data can be extrapolated to estimate performance under operational conditions. Modeling and simulation (M&S) can help quickly obtain initial data sets, perform sensitivity analyses, and drive additional testing questions. M&S can be a cost-effective method when there are limits on physical experimentation, though it should not replace operational testing (Marine Corps Operational Test & Evaluation Activity [MCOTEA], 2013). Simulation methods can be integrated with a test process, especially in developmental phases before a final assessment is made, and can be especially important in DT&E (*T&E Management Guide*, 2005).

DT&E can usually inform the types of experiments run in OT&E. DT&E plays a major role in evaluating a potential system and its ability to meet the capabilities requirements. In order to ensure that a proposed system meets the requirements, a detailed DT&E process is needed to test system capabilities, limitations, costs, and safety. The data carefully collected in these experiments provides a wealth of information that can be used to inform efficient OT&E exercises. Because the questions used to design an operational test plan are motived by the results of DT&E, there is a unique opportunity to leverage two-stage statistical methods to efficiently answer questions about whether the capability requirements have been met.

For example, DT&E can be used to screen potential tests that may be unnecessary in OT&E because it is deemed that certain configurations of a system are likely to have poor operational performance and no further effort should be wasted on these settings. While Design of Experiments (DOE) is often considered a critical part of OT&E, using it in DT&E can only enhance the types of experiments that could be run in OT&E. Ortiz and Harman

(2016) argue for the use of DOE in DT&E in addition to OT&E because randomization, replications, and blocking can be more easily implemented. Such experiments in DT&E can narrow the space of possible feasible configurations to test in OT&E. This is part of the “shift-left” mentality to do more analysis in earlier stages of development to save costs and improve results throughout the entire acquisitions process.

Because OT&E assesses the performance of a system under more realistic conditions, testing can be much more expensive and constrained. Thus, it is even more important to design a test plan that is able to obtain the best information possible given constraints on the overall testing budget across the two stages. Additionally, the research questions and decisions that need to be made may have changed as a result of DT&E. Understanding integrated testing and evaluation is critical to efficient implementation of modeling and simulation results (United States Marine Corps, 2010).

Confidence Intervals

Confidence intervals are commonly used to assess the risk associated with the system by evaluating mean performance. Here we give a brief summary of confidence intervals to define notation and introduce key parameters. A confidence interval is collected from *n* samples of system performance results to estimate the mean of the system *μ* using $\bar{X}$ as the centerpoint. The half-width on either side of the centerpoint defines the confidence interval

$$\left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{1}$$

where σ^2 is the variance estimate for the data. If the data is normally distributed and the variance is known, then the confidence interval can be estimated exactly using standard z-tables. If the variance of normal data is estimated, then *t*-tables are used. The two key parameters we study are the variance estimate, which is critical to understanding the risk associated with an estimate, and the sample size *n*, which is often controllable by the user. A larger variance estimate leads to a larger confidence interval. If the variance is underestimated, the confidence interval will be too narrow and there will be more certainty (than there should be) in the result. The sample size *n* is critical for estimating the variance, and it also determines the width of the confidence interval. More samples are better for reducing uncertainty in estimates, but often come at high cost in a T&E setting.

Choosing the Sample Size

Sequential methods for generating confidence intervals have been studied most recently in Singham and Schruben (2012) and Singham (2014). These methods increase the sample size until a confidence interval with a half-width smaller than some pre-specified level can be generated. They have traditionally been studied in the context of simulation models where large numbers of samples can be collected.

Suppose the estimate of the standard deviation is *s*, and we have some desired precision in our confidence interval *δ*, which is the half-width of the interval. Then, the sample size that guarantees (for independent and normally distributed data) that the confidence interval for *μ* has a half-width smaller than *δ* is

$$n \geq \frac{t_{n-1,\alpha}^2 s^2}{\delta^2} \tag{2}$$

and this can be used to choose the sample size. Johnson, Freeman, Hester, and Bell (2014) study sequential methods for estimating ballistic resistance of armor, and note that the methods used by the Department of Defense (DoD) have not changed recently. The methods can be simple to implement and do not require much statistical analysis, and the authors conduct simulation experiments to determine which tests are most effective at estimating different percentiles for the probability that the armor is perforated. Such tests are often used as part of Lot Acceptance Testing to determine whether a production item is acceptable.

Two-Stage Procedures

Two-stage procedures are often used instead of single-stage procedures because initial data collected in the first stage can be used to enhance the efficiency and quality of results in the second stage. A main example of this is using the first stage of an experiment to estimate the variance of the system. The variance is usually unknown ahead of time, yet it is a crucial part of estimating confidence intervals or other measures of performance. A poor variance estimate can lead to low validity of statistical results. Results from DT&E can be used to estimate the variance of the system, which in turn helps decide how many runs are needed in OT&E. For example, if the variance of the system is high, then more runs will be needed in OT&E to assess the feasibility of the system. If the variance of the system appears low, perhaps fewer runs will suffice.

Given a set of *n* independent and identically distributed (i.i.d.) samples of system performance estimates, then

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{3}$$

is used as the variance estimate. When the data is dependent and is normally distributed, we can quantify the dependence using autocorrelation with lag *h,* which is a measure of dependence between sample X_i and X_{i+h}. If the output data of a series has positive dependence, we hope that this dependence decreases over time as *h* increases, so that observations far apart are relatively independent. If the dependence between samples is positive, the variance estimate will be smaller than it really is. This means that the risk in the system will be underestimated, and we would proceed to OT&E with more certainty in performance than what actually exists.

Positive dependence between samples can exist for many reasons. For example, if a machine is not completely reset and recalibrated between samples, then the state left by the previous run can affect future runs. If the same operator tests the machine or weapon for each run, there may be correlation between outputs based on the habits or practices of the operator. In reality, there may be more variance in an operational setting because there will be many different people using the equipment. Thus, it is important to ensure independence between samples in the first stage. It may be useful to employ a confidence interval for the variance:

$$\left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{\alpha/2}}, \frac{(n-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] \tag{4}$$

where the chi-squared term is the relevant quantile of the chi-squared distribution with *n-1* degrees of freedom. This means that we can assess the uncertainty in the variance estimate based on the number of samples taken in the first stage, and inflate our estimate of the variance in the second stage using the upper confidence level of the variance estimate. Inflating the value of the variance estimate will encourage more samples to be taken in OT&E and will protect against the potential underestimation of risk resulting from a too-low variance estimate.

Ranking and Selection

Ranking and selection procedures attempt to determine the best system inputs when the system configurations are discrete options that can be listed. There is uncertainty ahead of time about the actual performance of the system, and the feasibility of the system to meet some constraints. Figure 1 shows the potential layout from the first stage of a ranking and selection experiment. The *x*-axis measures the feasibility of the system, while the *y*-axis measures the performance along the main objective or measure of effectiveness (MOE). The goal is to select the system with the best objective that is feasible. Based on the figure, it makes sense to invest more time in the second stage on the “Feasible, good objective” system and the “Infeasible, best objective” systems. It is possible the latter system may actually be feasible if we tested more, or it’s possible the former system may actually be the best system. In any case, it probably does not make sense to spend resources in the second stage on the “Feasible, poor objective” and the “Infeasible, poor objective” systems.

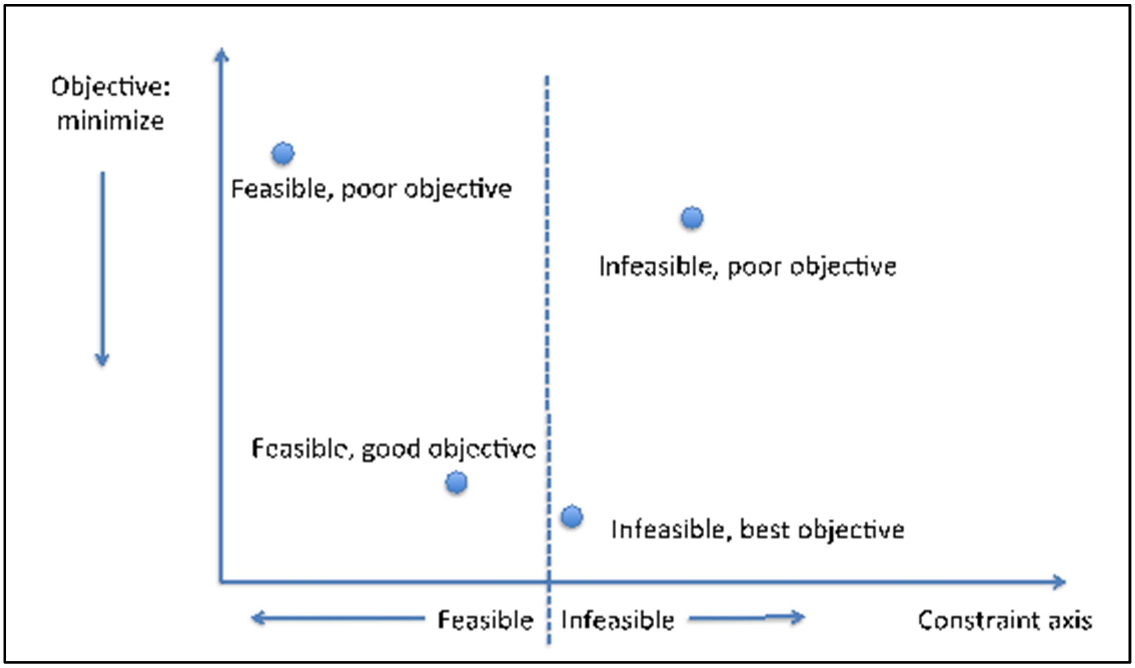


Figure 1. Comparison of System Configurations by Feasibility and Objective Function

Subset Selection/Screening

A number of subset selection procedures exist that screen out potential system configurations that are deemed suboptimal or infeasible. The first stage takes some initial number of samples from each system in the hopes of obtaining information that can be used for a more efficient second stage. In some cases, many system configurations can be eliminated from consideration in the second stage. This is something that occurs naturally in the transition between DT&E and OT&E; we do not usually bother to test options in OT&E that clearly did not work in DT&E.

One such subset selection procedure is Singham and Szechtman (2016), which uses information in the first stage to estimate the variance of the system and then allocate effort to the second stage accordingly. Systems with higher variance obtain a higher allocation of effort because they have more uncertainty. Similar methods can be used, as in Figure 1, to allocate sampling effort to systems close to the feasibility boundary, or close to optimality.

Then, in the second stage, a subset of the systems is chosen which is likely to contain the best systems with high probability.

A Two-Stage DT&E/OT&E Integrated Procedure

We now describe a two-stage statistical procedure that can be mapped to the stages of DT&E and OT&E. There are many different contexts to consider, but here we study the case where DT&E experiments allow for an arbitrary number of trials. For example, computer simulation experiments can often be used to test the potential readiness of a system, and it can be easy to run many replications.

The goal of the experiment is to determine which systems meet the requirements for performance, and, if more than one system meets the requirements, to determine which one is the best, or most cost-efficient, option. There are two main objectives of the first stage. The first is to screen out any system configurations that are highly likely to fail in OT&E, thus saving valuable experimentation resources. The second objective is to allocate resources to the remaining systems so that in OT&E the best system determination can be made. As in Figure 1, more resources would go to systems that are close to the feasibility boundary for meeting performance. Additionally, systems that display a high variance in the first stage would receive more samples in order to reduce their confidence intervals to make an operational suitability determination.

Next, we present the details of the two-stage statistical experiment. We run the first-stage experiments to estimate the mean and variance. These are used to calculate *p*-values, which are used to determine which systems can be eliminated from contention as worse than the threshold. Then, an inflated variance estimate is used to assign sample sizes to each system. This inflated estimate is used to account for potential model error resulting from the simulation setting being different from an operational setting.

1. The objective is to select the best alternative system that performs at least as well as the benchmark system, which determines the feasibility/capability requirements.
2. Develop DT&E experimentation parameters to answer objectives.
    a. For example, when analyzing performance of a sensor, two factors are (1) the coverage area of the sensor and (2) the location and number of sensors.
    b. Given the first stage is a simulation stage, we can run a large fixed number of replications of each system configuration to estimate the variance. However, to illustrate the effect of variance estimation in a limited budget, we run 30 replications of each configuration.
3. Run first-stage DT&E and analyze results.
    a. Estimate the mean $\bar{X}_i$ and variance $\hat{\sigma}_i^2$ for each system configuration *i*, including the benchmark system. Call the estimated mean for the benchmark $\bar{X}_0$ and, if the capabilities threshold for the benchmark is known, then its mean is fixed at *μ*.
    b. Reassess critical issues and specific objectives for the system, screen out factors and configurations if possible.
        i. Calculate *p*-values for each system for comparison to the system mean. Let *n* be the number of samples, and $F_{t_{n-1}}$ be the cumulative distribution function of the *t* distribution

with *n-1* degrees of freedom. If the benchmark is estimated then replace *μ* with $\bar{X}_0$ (see Singham and Szechtman, 2016, for an example of this type of calculation).

$$p_i = F_{t_{n-1}} \left(\frac{\overline{X}_i - \mu}{\hat{\sigma}_i / \sqrt{n}} \right) \tag{5}$$

ii. Use *p*-values to determine which systems to eliminate. These systems have a low probability of having performance that is better than the benchmark. For example, if

$$p_i \leq \alpha \tag{6}$$

then typically for $0 \le \alpha \le 0.1$,, eliminate the system from contention for having a mean performance level that is so small to be unlikely to be better than the benchmark *μ*. This will remove systems that have a small mean relative to *μ* while also having a relatively a small variance because we are fairly certain these systems will perform poorly.

c. Using confidence intervals for the sample variance, we can choose the upper confidence limit to deal with uncertainty associated with future OT&E experiments giving a conservative performance estimate.

$$\tilde{\sigma}_i^2 = \frac{(n-1)\hat{\sigma}_i^2}{\chi^2_{1-\alpha/2}} \tag{7}$$

d. Determine the budget allocation for the second stage based on first-stage results by comparing outcomes to the threshold objectives.

i. Calculate the sample size needed for each system to compare it to the threshold using properties of absolute and relative precision sampling as determined in Singham (2017).

$$n_i \geq \frac{t^2_{n-1,\alpha} \tilde{\sigma}_i^2}{|\overline{X}_i - \mu|^2} \tag{8}$$

ii. We need to do a similar calculation for the benchmark system if its true performance *μ* is not known. We decide a precision *δ >0*, which is the allowed deviation from *μ* that would be acceptable in a confidence interval estimate of the benchmark. Then, the second stage number of samples for the benchmark is

$$n_i \geq \frac{t^2_{n-1,\alpha} \tilde{\sigma}_i^2}{\delta^2} \tag{9}$$

iii. Rescale the sample sizes to be proportions for the second stage given a total budget *N*, and *S* total systems under testing.

$$\hat{n}_i = N \left(\frac{n_i}{\sum_{i=1}^{S} n_i} \right) \tag{10}$$

iv. If the unscaled n_i values are much too large for OT&E, then run n_i samples for system *i* in DT&E to obtain further information and repeat the screening, as in Step 3.b.ii, to remove additional systems that appear unlikely to beat the benchmark.

4. Run second-stage OT&E and analyze results.
   a. Run experiments on the potential subset using $\hat{n}_i$ sample sizes for each system *i*.
   b. Determine whether the requirements and objectives have been met by comparing the final results to the threshold. A similar *p*-value calculation to the one above can be used to determine if a system is significantly better or worse than the threshold.

What will most likely occur is that the first-stage experiment will determine a large number of samples n_i that will be needed to test each system. If these sample sizes are too large for OT&E, then we recommend running these experiments in DT&E to obtain as much information as possible and repeating Steps 3 and 4. The idea is that with enough samples, the difference $|\bar{X}_i - \mu|$ becomes large relative to $\hat{\sigma}_i/\sqrt{n_i}$ so that a clear determination can be made whether system *i* is better or worse than the benchmark *μ*. This can be used to screen out systems that are worse than the benchmark, and determine the allocation of effort toward systems better than the benchmark. Afterwards, if the number of samples is still too high for OT&E and there is a total budget *N* for samples, the rescaling can be done to allocate the budget towards systems that require more samples to make a determination.

In some cases, the T&E analyst may want to further reduce the subset from those that appear better than the benchmark for OT&E. For example, if seven out of 10 configurations are in the selected subset, the analyst may only choose the top three for consideration in OT&E to determine the best one.

Case Study—Unmanned Sensors for Intelligence Collection

To illustrate the procedure, we use a simulation experiment designed to test the performance of sensors for tracking targets such as pirates or smugglers. These sensors are designed to report information on potential targets of interest in large unpatrolled areas of water. Different sensors have different properties. For example, some have larger areas of coverage, while others may be more accurate and have a higher probability of detecting a target. The goal is to determine whether a particular sensor configuration can achieve the performance needed to be successful in finding targets, while balancing the cost and number of sensors to be purchased.

The simulation model has been built by the author and colleagues and is part of ongoing research being conducted at the Naval Postgraduate School. The full theoretical model details are available in Nunez, Singham, and Atkinson (2018). The model simulates numerous target paths given intelligence about the target's trajectory. Sensors can then be placed, and the number of target paths that are successfully observed can be recorded. Experiments can be run to determine a number of objectives, for example, which configuration is the best, or how often a particular setup successfully observes the target. We note that in this study, we do not consider whether physical specification requirements are met, but rather focus on whether the particular system can meet operational requirements.

Cheng (2016) studied different sensor configurations using this model to compare their performance. The benchmark given was Lynx multi-mode radar, which has a range of 80km (about 0.72 degrees) and an endurance of 48 hours. The Lynx radar system delivers high quality results but can be quite expensive (close to $7 million). Thus, we want to determine if we can obtain similar performance results using two cheaper unmanned sensors that may have smaller coverage areas. We use the sensor simulation model as the model to test the two-stage procedure. The model is flexible and allows for infinite input possibilities, and, as it is a computer simulation model, it is relatively inexpensive to run multiple replications to collect data.

Experimental Results

The experiment runs by simulating multiple potential target paths based on intelligence. Sensors are placed at the beginning of the run to attempt to locate the target as it passes through the area, and the simulation records the proportion of paths that intersect the sensor coverage areas. There will be variation each time the experiment is run due to randomness in the simulated paths. Thus, it is important to run multiple replications to estimate the potential error in the estimated probability of success.

We place sensors along the central expected path of the simulated target to obtain the maximum probability of success. Figure 2 shows the benchmark sensor placement for a target that is predicted to depart off the coast of South America towards the western coast of Mexico (red box). The blue heatmap shows the relative likelihood of the target's location given the intelligence at hour 25, with a higher probability in the middle. The Lynx sensor is positioned to anticipate observing the target at hour 50, but there is a high probability the target will not pass through the sensor and will remain undetected.

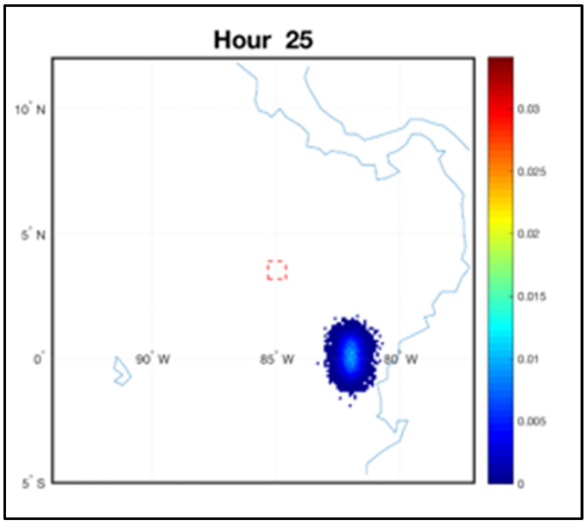


Figure 2. Benchmark Sensor Placement (Red Box) and Target Distribution (Blue Heatmap)

The alternative systems to the benchmark include those with two sensors with smaller coverage areas. We place the sensors to anticipate where the target will be at hours 35 and 70. While these sensors are smaller, there are two of them, so the second sensor may capture targets that remained undetected by the first sensor.

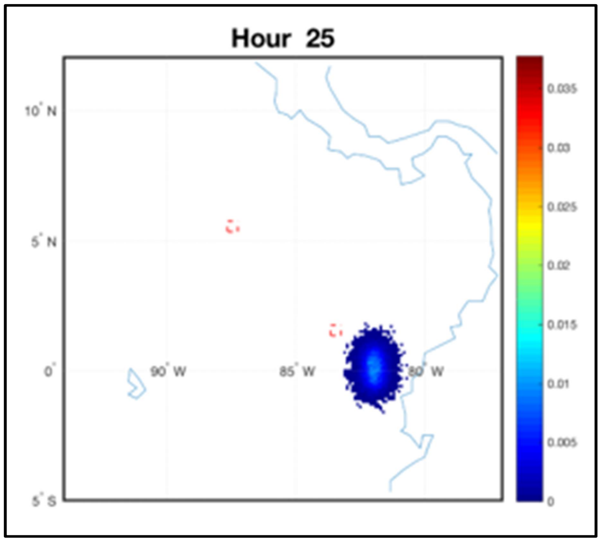


Figure 3. Dual Configuration: Alternative Sensor Placement With Two Smaller Sensors

We conduct first-stage experiments to compare different alternative configurations against the benchmark. The results of these experiments decide which systems have the potential to be better than the benchmark, and how to allocate second-stage experiments in OT&E. We note that sensors with smaller coverage areas are assumed to be cheaper and are preferred. A second-stage experiment could consist of more comprehensive simulation runs, or operational testing of the sensor in practice to see how it performs. Table 1 summarizes names of the system configurations, with the benchmark, dual sensor configurations, and their coverage widths.

Table 1. Names and Coverage of System Configurations

System Configuration	Coverage Width (each sensor, degrees)
Lynx (single) benchmark	0.72
Dual20	0.20
Dual30	0.30
Dual35	0.35
Dual37	0.37
Dual38	0.38
Dual39	0.39
Dual40	0.40
Dual50	0.50

All of the sensors in the dual configuration have much smaller coverage widths than the Lynx. We apply the algorithm to a series of first-stage experiments, as described previously, by running 30 replications of the experiment for each configuration and saving the mean and variance of the proportion of targets detected. Each replication simulates 200 target paths based on intelligence. We use these values to calculate p-values relative to the benchmark, and then eliminate systems who have *p*-values smaller than α=0.05, as these are unlikely to be better than the benchmark. For the remaining systems still in contention, we calculate the upper bound on σ^2 to determine the number of replications needed to distinguish the system from the benchmark mean. Table 2 summarizes the results of the experiment.

Table 2. First Stage Experiment Performance

System Configuration	First Stage Mean	*p*-value	Number of Samples	Proportion of Samples for Second Stage
Lynx (single) benchmark	0.1457	---	35	9%
Dual20	0.0408	0	---	---
Dual30	0.0892	0	---	---
Dual35	0.1155	0	---	---
Dual37	0.1273	0	---	---
Dual38	0.1407	.17	220	58%
Dual39	0.1543	.96	65	17%
Dual40	0.1542	.97	57	15%
Dual50	0.2235	1	2	1%

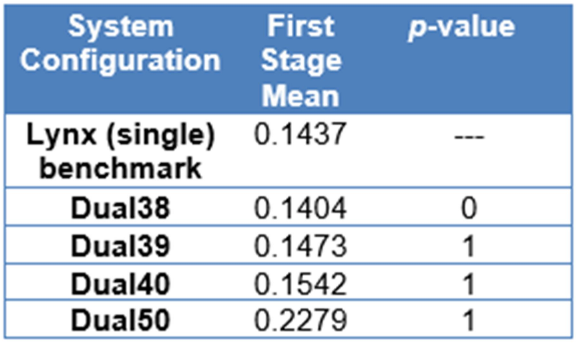
We require a precision of 1% on the estimate of the benchmark, so the allowable deviation in the estimated performance of the benchmark is 1%. The systems with two sensors with small coverage areas (Dual20, Dual30, Dual35, Dual37) all have estimated performance significantly below that of the benchmark, so the *p*-value is 0. We can eliminate these systems from consideration in the second stage. It is apparent that Dual50 has the best performance by far, with Dual38, Dual39, and Dual40 having performance close to that of the Lynx single sensor system. Depending on the requirements, we may want to choose the sensors with the smallest coverage width if they are cheaper.

We use the algorithm to calculate the number of samples needed in the second stage for the remaining systems and the benchmark. The Lynx system requires 35 samples to estimate the mean performance down to 1% absolute error. The Dual50 system only requires 2 samples, mainly because its performance is much higher than the benchmark, so little additional testing is needed to distinguish it as an improvement. The Dual38 system requires 220 samples because its performance is closest to that of the benchmark, so many more samples are required to distinguish whether or not it is better. Dual39 and Dual40 require 65 and 57 samples, respectively, to ensure they are better than the benchmark.

The last column shows the percentage of effort needed for each system. If the second stage cannot complete the recommended sampling effort because of cost or operational constraints, the last column shows the relative effort that should be expended on each system, with 58% of the effort going to Dual38. At the end of the second stage, we hypothesize that Dual39 is the “cheapest” system that has performance at least as good as the benchmark, where the smaller coverage area sensors are cheaper. However, we must still expend significant effort on Dual38 because it could be better or indistinguishable from the benchmark.

We conduct a second-stage experiment, which is meant to represent a more expensive operational setting but still involves a simulated model. Each replication now simulates 20,000 independent target paths (instead of 200 in the first stage), resulting in a more accurate estimate. In reality, the second-stage experiments would be in an operational setting where real information could be obtained.

Table 3. Second Stage Experiment Performance



System Configuration	First Stage Mean	*p*-value
Lynx (single) benchmark	0.1437	---
Dual38	0.1404	0
Dual39	0.1473	1
Dual40	0.1542	1
Dual50	0.2279	1

The second-stage results in Table 3 show clearly that Dual38 does not perform as well as the benchmark, while Dual39, Dual40, Dual50 are superior to the benchmark. Thus, the conclusion is that Dual39 is the cheapest system that performs at least as well as the benchmark, meaning two sensors with a coverage width of 0.39 would perform at least as well as one sensor with a coverage width of 0.72. However, the analyst could still choose Dual38 if she or he felt it was close enough to meeting the requirements. We note that the first stage required 270=9x30 total replications, while the second stage required 379 total replications. By eliminating some systems after the first stage and reallocating effort, we are able to focus effort on obtaining the best system. This saves effort over continuing to employ equal allocation over all systems in the second stage.

Conclusion

We present a two-stage statistical method that can be used to link experimental parameters in DT&E and OT&E experiments. The first-stage experiments can be used in DT&E to estimate the performance of different systems. These results can be analyzed to determine which system configurations to test in OT&E and how to allocate effort in the second stage. Typically, more effort should be allocated towards systems with high variance or those close to the feasibility boundary or capabilities requirement, which can be determined by a benchmark or other metric. We apply the algorithm to a model designed to compare different sensor configurations.

References

Allen, C. L., Jr. (2010). The role of simulation in test and evaluation. *ITEA Journal, 31*, 378–383.

Cheng, C. C. (2016). *A Brownian bridge movement model to track mobile targets* (Master's thesis). Monterey, CA: Naval Postgraduate School.

Chick, S. E., & Inoue, K. (2001). New two-stage and sequential procedures for selecting the best simulated system. *Operations Research, 49*(5), 732–743.

Giadrosich, D. L. (1995). *Operations research analysis in test and evaluation*. Reston, VA: American Institute of Aeronautics and Astronautics.

Goldsman, D., & Nelson, B. L. (1998). Statistical screening, selection, and multiple comparison procedures in computer simulation. In *Proceedings of the Winter Simulation Conference, 1*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Hill, R. (2017). The test and evaluation workforce and a base of sand issue. *ITEA Journal of Test and Evaluation, 38*(2).

Johnson, T. H., Freeman, L., Hester, J., & Bell, J. L. (2014). A comparison of ballistic resistance testing techniques in the Department of Defense. *IEEE Access, 2*, 1442–1455.

Marine Corps Operational Test & Evaluation Activity (MCOTEA). (2013). *MCOTEA operational test and evaluation manua*l. Quantico, VA: United States Marine Corps.

Nunez, J. A., Singham, D. I., & Atkinson, M. P. (2018). *A particle filter approach to estimating target location using Brownian bridges*. Manuscript submitted for publication.

Ortiz, F., & Harman, M. (2016). DOE in DT: The place to be*! ITEA Journal of Test and Evaluation, 37,* 241–245.

Singham, D. I., & Schruben, L. W. (2012). Finite-sample performance of absolute precision stopping rules. *INFORMS Journal on Computing, 24*(4), 624–635.

Singham, D. I. (2014). Selecting stopping rules for confidence interval procedures*. ACM Transactions on Modeling and Computer Simulation (TOMACS), 24*(3), 18.

Singham, D. I. (2017). Decision-based metrics for test and evaluation experiments. In *Proceedings of the 14th Annual Acquisition Research Symposium*. Monterey, CA: Naval Postgraduate School.

Singham, D. I., & Szechtman, R. (2016). Multiple comparisons with a standard using false discovery rates. In *Proceedings of the 2016 Winter Simulation Conference*. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Test and evaluation management guide (5th ed.). (2005). Fort Belvoir, VA: Defense Acquisition University Press.

United States Marine Corps. (2010, May 6). *U.S. Marine Corps integrated test and evaluation handbook* [Memorandum]. Washington, DC: Author.

ACQUISITION RESEARCH PROGRAM
GRADUATE SCHOOL OF BUSINESS & PUBLIC POLICY
NAVAL POSTGRADUATE SCHOOL
555 DYER ROAD, INGERSOLL HALL
MONTEREY, CA 93943

www.acquisitionresearch.net