Quantifying STAT Impact in DoD Test Design

Authored by: Sarah Burke, PhD
Michael Harman
Francisco Ortiz, PhD
Lenny Truett, PhD
Aaron Ramert
Bill Rowell, PhD

15 February 2017

Revised 27 August 2018

The goal of the STAT COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.afit.edu/STAT.

STAT Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Revision 1, 27 Aug 2018, Formatting and minor typographical/grammatical edits.

Executive Summary

Department of Defense (DoD) leadership seeks to improve the quality and efficiency of test and evaluation through the use of scientific test and analysis techniques (STAT). Measuring the impact of STAT requires the calculation of a return on investment (ROI) metric to understand the value of incorporating STAT into test and evaluation. Since ROI measures have not yet been defined for STAT in DoD, this paper proposes a solution as a starting point. Our goal is to start a discussion on how to improve these metrics and generate other relevant methods to effectively communicate the value of STAT. The possible metrics discussed are related to cost, schedule, and knowledge and their associated challenges. Knowledge improvement was the resulting metric selected and defined and is presented here with a practical example showing how to monetize the additional knowledge gained through a STAT-based design. Although STAT can be applied to many aspects of test and evaluation (test design, automated software testing, performance testing, etc.), this paper focuses on its application to test design.

Keywords: return on investment, quantification, monetize, test design

Introduction

Department of Defense (DoD) leadership seeks to improve efficiency (Kendall, 2016) and wants to understand the rationale used to select test points (OSD, 2015). Scientific test and analysis techniques (STAT) is the use of scientific methods in developing rigorous, defensible test and analysis plans. STAT aims to improve the planning, execution, analysis, and reporting of testing. An assessment of the return on investment (ROI) is important to understand the value and impact of applying STAT. Better understanding the value of STAT will also encourage increased use of STAT throughout the DoD.

While we focus on applications of STAT in the DoD, this is not a problem unique to DoD. Although quantifying the value of STAT is a common issue in industry, healthcare, and government, our research could find little to no information on how to assess the value of STAT, for example, by monetizing the value of a designed experiment.

Determining an ROI requires quantification of relevant measures. Phillips (2003) discusses a step-by-step process on how to calculate the ROI of a training program. Many of the methods Phillips discusses on converting metrics to cost simply don't exist or don't apply in the DoD environment. For example, Phillips suggests using historical records to estimate the value of a unit of improvement. While these historical records are commonly available in companies, they are not in the DoD. For STAT in DoD, ROI measures have not yet been defined; this paper proposes a possible solution as a starting point. Our goal is to start a discussion on how to improve these metrics and generate other relevant methods.

Motivation

ROI requires some basic numbers for its calculation: an investment cost and an output value ("Return on Investment ROI"). The ROI can be expressed as a difference between output and investment values or as a ratio. A positive difference, or ratio greater than 1.0, indicates improvement, with the larger values/ratios indicating a larger impact. ROI methods rely on the assumption that anything can be measured and, necessarily, the ability to quantify those measures (Phillips, 2003). The traditional measures of performance improvement are quality, cost, and time. Test and evaluation (T&E) process improvement assessment can follow a similar method; however, this is particularly difficult in DoD where there is a lack of granular data relating the myriad details of testing to overall test costs. A cost history certainly exists for each test effort, but it tends to be bottom line information. In industry, if historical records are not available for a given metric, data is frequently collected via surveys to measure impact. Many of the measures in DoD T&E cannot be measured through surveys and we don't produce items on a production line making it difficult to measure impacts from implementing STAT.

Possible Metrics

To create a STAT ROI metric, we consider the following:

- Cost savings
- Schedule savings
- Knowledge gain

Improvements directly correlated with STAT in any of these areas would indicate a positive STAT ROI. However, there are different challenges in using each of these metrics. Primarily, ROI assumes there is a current state and an improved state that can be compared. This assumption implies that the current methods are not employing STAT effectively so that STAT-based design improvements can be made and the appropriate ROI metrics can be computed.

Cost savings, as a metric, is easily defined (planned vs actual), but its computation can be highly subjective and non-linear, for example, due to discontinuities. For instance, the set-up cost for a range visit might be fixed; going for a week or a month incurs the same setup cost, despite different test duration costs. Also, cost savings imply some money is left in the bank. While theoretically true, any savings due to reduced test costs would most likely be moved to fund other risk areas resulting in no bottom line savings. Also, early test and evaluation master plan (TEMP) resource planning may not actually generate a line-item list of tests and costs. Without a discrete cost breakout, it would be difficult to state what the "before STAT" planned cost was so that an "after STAT" cost can be calculated and compared.

Similarly, schedule savings can be nonlinear and a savings of small amounts over several tests (e.g. days saved over several months) may be lost in the noise created by weather delays, range issues, and other common occurrences. Large STAT-generated savings would have to be factored into the TEMP resources

at the beginning and, unless STAT impacts every test, are unlikely to make a clearly discernable difference at the macro planning level.

The gain in knowledge or information resulting from STAT is not what is typically expected in an evaluation of the impact of STAT. We believe this is one of the critical advantages of STAT however difficult it may be to quantify. Therefore, we focused on this metric as a path forward to quantify the impact of STAT.

Defining a Knowledge Metric

Testing is used to inform a decision; however, quantifying the amount of knowledge or information used to inform a specific decision is hard to determine. Knowledge improvement as a measurement is equally difficult to assess. Any ROI measure must assess knowledge before and after STAT implementation. Since not every test will employ STAT, this metric is best applied to a single design. For an overall assessment, metrics generated for several designs can be incorporated into an overall STAT assessment.

Options and Thought Process

The various ideas under consideration were the number of test points, location of test points, risk of drawing the wrong conclusion, and precision in assessing the requirement. The number of test points alone is not a good measure because it does not convey anything about the quality of the information obtained through them and follows the dubious "more is better" theory. The location of the points in the design space may be a better indicator but can be subjective and biased. This is especially true if there are regions in the space that are not covered (but should be) and, therefore, cannot be assessed for quality. Decision-making risk is a good metric as long as the generation of the value is objective and clearly understood. Another method would be to compare a test design before and after STAT is applied. This method does imply that there is a "before" design to compare to the "after" design, a situation we expect to become less common with the incorporation of STAT early on in the test process.

Resulting Metric

Using the "before and after" idea, we created a metric based on some of the quantitative design measures used in design of experiments. The Weighted Ratio Product Knowledge Metric (WRPKM) is a product of three ratios: number of points, number of model terms, and prediction variance:

$$WRPKM = \left(\frac{\#Points_{Orig}}{\#Points_{DOE}}\right)^{w_1} \times \left(\frac{\#ModelTerms_{DOE}}{\#ModelTerms_{Orig}}\right)^{w_2} \times \left(\frac{PredVar_{Orig}}{PredVar_{DOE}}\right)^{w_3} \quad (1)$$

The placement of the values (Orig = "before", design of experiments (DOE) = "after") in either the numerator or denominator is done to create a ratio where a value greater than 1.0 indicates the STAT design has increased knowledge.

- The Points ratio puts the DOE value in the denominator because a design with fewer points is more efficient and desirable.

- The Model Terms ratio compares the effective model size achievable with the given design points. Because a design with a larger number of well-chosen model terms should provide more knowledge and information, this ratio has the DOE value in the numerator.
- The Prediction Variance ratio places the DOE value in the denominator since a well-scoped design should produce a lower prediction variance than other design types. A smaller DOE prediction variance is desirable because the information provided is more precise and, in the denominator, will drive the ratio up. The value used is either at a pre-determined fraction of the design space (e.g. 50% or 90%) or the average over the entire design space. This choice should be documented and be consistent.
- The weights w_1, w_2, w_3 should be chosen such that $\sum_{i=1}^{c} w_i = 1, w_i \geq 0$ for all $i,\ and\ c = 3$.

We also formulated an alternate Weighted Ratio Sum Knowledge Metric (WRSKM):

$$WRSKM = w_1\left(\frac{\#Points_{Orig}}{\#Points_{DOE}}\right) + w_2\left(\frac{\#ModelTerms_{DOE}}{\#ModelTerms_{Orig}}\right) + w_3\left(\frac{PredVar_{Orig}}{PredVar_{DOE}}\right) \quad (2)$$

Weighting schemes ("*w*" terms) were introduced by Derringer and Suich (1980) as methods to measure a desirability score over various criteria in optimization problems. The choice between the additive or multiplicative form depends on the goals or needs of a given scenario. The multiplicative form penalizes poor performance more than the additive form. The additive penalizes a poor score less so that a good score can outweigh a poor one. While straightforward mathematically, we recommend documenting the rationale for the choice of metrics and any weighting scheme that is used.

A practical and ready to use spreadsheet (Quantifying STAT Impact in DoD Test Design - Tool) is available at https://www.afit.edu/STAT (under TOOLS) or via email at COE@AFIT.edu.

Implementation

Can Other Ratios Be Added?

The proposed metric can be modified (ratios added or removed) if other ratios are desired. Some additional candidates are:

- Confidence and Power: Type I and Type II risks
- Design Efficiency (Myers, 2016)
    - D-Efficiency: a measure to compare the quality of the model parameter estimates of designs with different sizes. Relative number of runs (in percent) that would be required by an equivalently sized orthogonal design to achieve the same determinant |X'X|.
    - A-Efficiency: This criterion evaluates how well the model parameters are estimated.
    - G-Efficiency: Evaluates a design for the maximum value of the prediction variance in the test space.

  - I-Efficiency: Evaluates a design based on the average value of the prediction variance in the test space.
  - V-Efficiency: Evaluates a design based on the average prediction variance over a specific set of points in the test space.
- $E[s^2]$ criterion
  - Minimizing this criterion is equivalent to minimizing the sum of squared off-diagonal elements of the correlation matrix (Myers, 2016). This is applied to supersaturated designs where the degrees of freedom for all main effects and the intercept term exceed the total number of distinct factor-level combinations (Gupta, 2011).
- Trace (AA')
  - This metric compares total bias in a design. This can be applied to computer simulations and screening designs (Myers, 2016)

These ratios could be added in a manner similar to the others. These were considered, but excluded for several reasons. We did not want to overly complicate the ratios by adding too many terms into the metric. If additional terms are included, we recommend documenting the rationale for why it is necessary and how it is calculated to create a single value (e.g. Power: average power of all main effects).

What if Some Ratios are Not Used?

If any of the three ratios is not required, then the specific ratio is set to 1.0 (or the corresponding weight to 0)

What if a Term is Not Defined?

The "original" test design may not have good statistical properties and, therefore, the determination of prediction variance or number of model terms may not be straightforward. Given this situation, it is easily argued that the DOE improves the amount of information because the original design was incapable of providing it. For these cases, we determined that the ratio in question would be assigned a fixed (albeit arbitrary) value of 5. In effect, this value states that the DOE provides five times the information over the original design. This value is shown clearly in the spreadsheet example in Appendix B.

What if the Final Metric Value is Less Than 1.0?

An overall value of less than one seemingly indicates STAT did not improve the amount of gained knowledge. This situation will typically occur if the ***Points*** ratio overcomes the ***ModelTerms*** and ***PredVar*** ratios. This is possible if the original test design is too small to affect the analytical outcome desired, requiring the DOE to add more points. The other way to look at this situation is that the original testing was not effectively sized initially and required more points to generate the required knowledge. Despite the seeming lack of improvement from this metric, it indicates the STAT process more effectively scopes the proper design size.

Document All Assumptions and Intangibles

Listing all relevant assumptions is critical for anyone evaluating the metric to be able to recreate the values and understand the logic. This documentation keeps the process transparent and repeatable and may, after sufficient practice, facilitate better definition of the process.

Intangible items should also be included. These items might include something the team considered, but could not quantify, such as an improvement to range operations imparted by the new design or improved leadership faith in the subsequent results. Similarly, complex design types like split-plot conditions certainly have an impact on knowledge gained. Simply comparing design metrics may not tell the complete story if the addition of a split-plot structure was required to correct and improve the quality of the analysis being done.

Using the Metric to Monetize the Knowledge Improvement

The knowledge statistic states, as a ratio, how much more information is gained using a STAT design. Describing this additional information as a cost savings (more information for the same cost) is the next step. Multiplying the ratio by the test cost results in an equivalent cost to gain this information using the previous test design type or size. Note that we do not imply that this is an exact number. Rather, this value is a probable order of magnitude cost that would have been required to gain this improved level of information had testing followed the original test design.

$$Ratio \times Test\ Cost = Information\ Value\ (Should\ Cost) \tag{3}$$

$$Example\text{:}\ \ 1.7 \times \$100{,}000 = \$170{,}000$$

Calculating the difference from the planned test cost provides a value of cost savings.

$$Information\ Value\ (Should\ Cost) - Test\ Cost\ (Planned) = Cost\ Savings\ (Avoided) \tag{4}$$

$$Example\text{:}\ \ \$170{,}000 - \$100{,}000 = \$70{,}000$$

The savings are an avoided cost because the money was not actually spent to obtain this additional information. One will notice this value can be calculated before any testing is started, given an existing test design, a new STAT design, and an original test cost. Furthermore, any reduction in the number of test points due to the new design can be used to revise the planned test cost down, which results in actual cost savings. The details of all the calculations (weighted/unweighted sum or product ratio, information value multiplication, and savings calculations) can be seen clearly in the spreadsheet.

Additional Metrics

Similarly created ratios for reliability and combinatoric designs produced knowledge metrics which are contained in Appendix A and included in the spreadsheet referenced in Appendix B.

Future Work

As stated, this paper was created to start a discussion on the topic for people facing similar requirements to quantify their STAT impacts. Aside from the metric, a major difficulty implementing this process is determining the specific test cost. Also, evaluating the previous design is possible but if new processes, procedures, or facilities are introduced, there may be no previous design to compare against.

The next step might be to start a trial period for measurement and employment. Using actual examples will bring to light any flaws in the logic and will help mature the process and metric. Using actual data will also help generate a list of relevant applications for these metrics and potentially generate others. The modification of the metric to include other terms might be required and a closer look at the reliability and combinatoric metrics may be in order. Previous discussion highlighted a potential flaw in the reliability metric (shown in Appendix A) if the number of test points grows in a STAT design and drives the metric below 1. A possible solution might be to scale the ratios to be between 0 and 1 so no ratio can overcome another one simply due to different measurement scales. This means that the final metric would also be bounded between 0 and 1, potentially creating an issue with monetizing the value. A trial run with the metric during actual planning and testing will help discern the size of this issue and may shed light on a solution.

Additionally, many test planning activities are managed by competency-aligned organizations (e.g. electronics, munitions, tracked vehicles group, etc.) that support numerous programs and provide consistency and expertise in a focused area. These competency-specific test designs may have room for improvement through the application of STAT principles. Introduction of STAT expertise and methods to the competencies, not just at the program level, will serve to improve methods for everyone the competency serves. Resident STAT experts at these facilities have access to all the subject matter expertise along with being an inside team member with the ability to audit current test procedures and to suggest practical improvements.

References

"Automated Combinatorial Testing for Software (ACTS)", *Computer Security Resource Center (CSRC) National Institute for Standards and Technology (NIST)*, 24 May 2016, csrc.nist.gov/groups/SNS/acts/index.html. Accessed 31 January 2017.

Derringer, George, & Suich, Ronald., "Simultaneous Optimization of Several Response Variables." *Journal of Quality Technology,* vol. 12, no. 4, 1980, pp. 214-219.

Gupta, V. K., et al. "About Supersaturated Design", *Indian Agricultural Statistics Research Institute (IASRI),* 18 Oct 2011, www.iasri.res.in/design/Supersaturated_Design/SSD/Supersaturated.html. Accessed 31 January 2017.

Kendall, Frank, "Better Buying Power Principles: What Are They?", Defense Acquisition, Technology and Logistics, A Publication of the Defense Acquisition University, January – February 2016, http://dau.dodlive.mil/files/2015/12/DATL_JanFeb_2016.pdf. Accessed 31 January 2017.

Myers, Raymond H., et al. *Response Surface Methodology*. Wiley, 2016.

Office of the Secretary of Defense (OSD), "Operation of the Defense Acquisition System," DoD Instruction 5000.02, Paragraph 10. a., *Executive Services Directorate*, 7 January 2015, www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002_dodi_2015.pdf?ver=2017-08-11-170656-430. Accessed 31 January 2017.

Phillips, Jack J. *Return on Investment in Training and Performance Improvement Programs*. London, Routledge, 2003.

"Return on Investment ROI", *Building the Business Case*. Solution Matrix Ltd, Publisher, www.business-case-analysis.com/return-on-investment.html. Accessed 31 January 2017.

Truett, Lenny, "Using Operating Characteristic (OC) Curves to Balance Cost and Risk-Best Practice." Scientific Test and Analysis Techniques Center of Excellence (STAT COE), 2013.

Appendix A: Metrics

DOE Knowledge Metric

This metric is the focus of the paper.

Version 1: weighted ratio product knowledge metric (WRPKM)

$$WRPKM = \left(\frac{\#Points_{Orig}}{\#Points_{DOE}}\right)^{w_1} \times \left(\frac{\#ModelTerms_{DOE}}{\#ModelTerms_{Orig}}\right)^{w_2} \times \left(\frac{PredVar_{Orig}}{PredVar_{DOE}}\right)^{w_3}$$

Version 2: weighted ratio sum knowledge metric (WRSKM)

$$WRSKM = w_1\left(\frac{\#Points_{Orig}}{\#Points_{DOE}}\right) + w_2\left(\frac{\#ModelTerms_{DOE}}{\#ModelTerms_{Orig}}\right) + w_3\left(\frac{PredVar_{Orig}}{PredVar_{DOE}}\right)$$

Reliability Knowledge Metric

This metric is applied using a reliability test time calculated using an operational characteristic (OC) curve (Truett, 2013). The reliability knowledge metric (RKM) is:

$$RKM = \left(\frac{Power_{STAT}}{Power_{Orig}}\right)^{w_1} \times \left(\frac{Conf_{STAT}}{Conf_{Orig}}\right)^{w_2} \times \left(\frac{\#Points_{Orig}}{\#Points_{STAT}}\right)^{w_3}$$

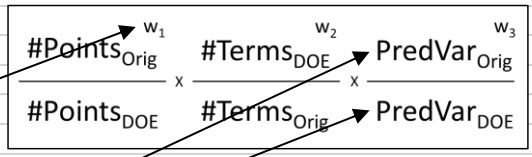
Combinatoric Knowledge Metric

The combinatoric knowledge metric (CKM) is applied using a combinatorial design type where *T-Way* represents the highest number of interactions/combinations supported by the design ("Automated Combinatorial Testing for Software", 2016).

$$CKM = \left(\frac{\#Points_{Orig}}{\#Points_{STAT}}\right)^{w_1} \times \left(\frac{T-Way_{STAT}}{T-Way_{Orig}}\right)^{w_2} \times \left(\frac{\#Factors_{STAT}}{\#Factors_{Orig}}\right)^{w_3}$$

Appendix B: Practical Spreadsheet Example

Type II Cost Savings			
Ratio for "0" before values		5	*When a "Before" value is zero, use this ratio instead (constant for all COE)*
DOE Value Metric			
# Points Before		100	
# Points After		50	
SubRatio		2.0	
Weight		0.3	
# Terms Before		2	
# Terms After		12	
SubRatio		6.0	
Weight		0.3	
Pred Var Before		1	
Pred Var After		0.7	
SubRatio		1.4	
Weight		0.4	
Unweighted Final Ratio		**17.1**	*Ratio of information gained with STAT compared to original method*
Weighted Additive Ratio		**3.0**	
Weighted Product Ratio		**2.4**	
Test Cost	$	1,000,000	
Unweighted Should Cost	$	17,142,857	*Cost to gather this information using original method*
Weighted Additive Should Cost	$	2,971,429	
Weighted Product Should Cost	$	2,430,609	
Unweighted Cost Savings	$	16,142,857	*Costs not spent for information actually obtained*
Weighted Additive Cost Savings	$	1,971,429	
Weighted Product Cost Savings	$	1,430,609	



- Blue cells indicate entered data
- Ratio for "0" before values: ratio used if the "Orig" value is undefined
- SubRatio: specific ratio calculation (e.g. # $Points_{Orig}$/#$Points_{DOE}$)
- Weight: Sum of all three must equal 1.0. First two weights are entered and the third is the difference W_3=1.0-W_1-W_2.
- Unweighted Final Ratio: Product with no weighting
- Weighted Additive Ratio: Sum of three weighted ratio terms
- Weighted Product Ratio: Product of three weighted ratio terms
- Test Cost: Planned or actual cost to execute this test
- Should Cost: Ratio x Test Cost (cost to gather this information using original method)
- Cost Savings: Should Cost – Test Cost (costs not spent for information actually obtained)

Similar data entry sheets for the reliability and combinatoric metrics are contained in the spreadsheet (Quantifying STAT Impact in DoD Test Design - Tool) available at www.AFIT.edu/STAT (under TOOLS) or via email at COE@AFIT.edu.