Screening the important factors in Supportability Test and Evaluation activities for Ships

Matthew Wernas
Royal Australian Navy Test, Evaluation & Acceptance Authority
matthew.wernas@defence.gov.au

Keith Joiner
University of New South Wales
k.joiner@adfa.edu.au

ABSTRACT

Research was conducted on historical data from supportability test and evaluation of Royal Australian Navy ships during their operational acceptance. The research used Design of Experiments test analysis techniques to identify which logistic and operational input factors have the greatest effect on the ability to complete planned maintenance of Navy ships. The 279 planned maintenance assessments used for the analysis were from Her Majesty’s Australian Ships Canberra and Adelaide as part of three major introductory service milestones. There were 25 supportability factors assessed across each of the supportability assessments. A nine-point Cooper-Harper rating scale was adapted to code the output response.

Nine supportability factors and eight two-factor interactions were found to be significant, representing a significant focusing from all of the combinations possible in the current testing, although not a valid model from which to predict outcomes. Pareto charts of the regression coefficients found that the factors of ‘qualifications of maintainers to do the maintenance tasks’ and ‘logic flow of maintenance instructions’, plus three of these factors’ interactions accounted for 57 percent of overall effect in supportability ratings. These key factors and their important interactions can be verified during ship developmental and acceptance testing using maintainability demonstrations involving representative Navy maintenance personnel, as first proposed by Blanchard’s influential logistics texts in 1974. This means Navy operational supportability assessments can be better prepared and de-risked using contracted maintainability demonstrations in ship development and design acceptance.

The research recommends the supportability assessment questionnaire be revised using lessons learned from this research analysis and checked for resolution and repeatability using a similar formal measurement system analysis during the upcoming operational testing on the new Hobart Class ships.

Keywords: Supportability Test and Evaluation, Maintainability, Design of Experiments, Screening Factors, Maintenance Demonstrations, Naval Ships
INTRODUCTION

In 2017 test staff from the Royal Australian Navy Test, Evaluation and Acceptance Agency (RANTEAA) undertook a course in advanced test and evaluation techniques by the University of New South Wales (UNSW) in Canberra as part of a Master of Systems Engineering by coursework, and then applied these techniques to the test design and test analysis of ship assessments. The techniques taught on the UNSW course are part of test competencies that have been mandatory for all acquisition and operational staff with test functions in the U.S. Department of Defense since 2009 (Defense Acquisition University, 2015). Moreover, the Office of Secretary of Defense requires all U.S. DoD test and evaluation plans to objectively use these highly efficient and rigorous techniques in the test design and test analysis of their new and legacy systems (Ahner, 2016; Chu, 2016), such that Australian Defence are new to using these techniques (Joiner, McAuliffe & Kiemele, 2016). The techniques are primarily based on Design of Experiments (DOE) theory and software packages for optimal multi-factor, multi-response test design and analysis (Reagan & Kiemele, 2008), but also include combinatorial techniques for mixed levels and efficient representative sampling methods for validation testing.

Research by the authors aimed to use the DOE analysis techniques to identify which logistic and operational input factors have the greatest effect on the ability to complete planned maintenance (PM) for Royal Australian Navy (RAN) ships. The information sought was to help RANTEAA: (1) provide better feedback to ship Project Directors and discipline leads in the RAN for logistics, safety, engineering and so forth, (2) more readily identify areas to focus improvements for ships being introduced to service, and (3) focus inspection and test effort for ship support systems. The planned maintenance assessments had been conducted four times prior to this analysis, with three of those four being on Her Majesty’s Australian Ships (HMAS) Canberra and Adelaide during the three main introduction into service milestones. The analyses followed the guidelines for researching historical qualitative data set out in Reagan and Kiemele (2008, Section 7.16).

RANTEAA’s processes for assessing the suitability of a ship during operational test and evaluation examines many activities by the delivery agency or supporting groups that need to have occurred in designing and preparing for a ship’s organic planned maintenance routines. For example, the suitability process evaluates for planned maintenance:

- the skill level required and the category of maintenance staff;
- the number of maintenance personnel required;
- the location of necessary equipment;
- any safety considerations;
- the adequacy of maintenance training;
- if necessary maintenance isolations and facilities have been identified, and
- adequate identification, acquisition, codification and delivery to the unit of necessary resources such as: tools, support and test equipment, spare parts, consumables and personal protective equipment (PPE) to perform the planned maintenance.

The process flow diagram for the planning and conduct of organic planned maintenance is provided in Figure 1. The current assessment process uses a set of 29 questions each being two to four levels and each question having room for comments. In gathering test data, RANTEAA test staff either witness the planned maintenance routines being conducted or they conduct interviews with maintenance staff after the test activity.
As part of DOE preparation, a cause and effect analysis was performed (Figure 2) to confirm adequate capture of the causal elements that influence the likelihood of completing planned maintenance on the ship’s system using representative personnel who are following approved procedures in the representative environment. Ideally the support system elements, including training, and operational authority procedures would be constants (labelled C), the ship’s routine and the environment would be noise (labelled N), and the remaining factors would be the experimental inputs (labelled X) to the assessment process. However due to the immaturity of the support system elements and operational procedures when introducing new capability the constant factors were considered as experimental inputs for the analyses.

The use of historical data limited the factors available for the DOE processes, for example, where only the proposed time to complete the maintenance had been captured and not the actual time taken. The maintenance management system used by the ship class is called Asset Management Planning System (AMPS) and the maintenance times as recorded within AMPS are typically the automatically pre-entered time as in the PM routine; not the actual times as conducted by the maintainer. AMPS also does not separate the duration of the routine into administration and logistics time and maintenance time. These measurements should be included in future test events.

The information collected within the RANTEAA questionnaires does not include weather, ship’s routine, a detailed time breakdown, or the assessor, as indicated in the cause and effect (Figure 2), so these were included as noise for this analysis. In future test events a preparatory measurement system analysis (MSA) should be conducted in accordance with the full rigour of recommended DOE practices (Reagan & Kiermele, 2008).
The 29 questions asked within the assessment questionnaire are all possible effects on an overall assessment outcome for screening key attributes in the process. Of the initial model’s 29 factors, 14 were two-level factors, 13 were three-level factors, and two were four-level factors. Every factor is a qualitative factor that is represented numerically and that has stepped input changes that ought to be categorical, but for simplicity in the screening regression analyses, were treated as quantitative. This was also a necessary limitation of the software used on the UNSW course. The initial input-output model is shown in Figure 3.
ANALYSIS PREPARATION AND FURTHER LIMITATIONS

During the operational test periods hundreds of planned maintenance routines were assessed. Each routine was assessed using a standard questionnaire, the answers being the information for each run for the screening and validation to be conducted. 122 planned maintenance routines from the initial test period on the first ship and 157 planned maintenance routines from the next test period on a second ship were assessed as valid and utilised for the analysis. Each planned maintenance routine assessment was considered as an individual run with no repetitions due to the variance in maintainers and assessment staff. The result of this is the screening model has no capability to determine the factors affecting variance. Many of the questions within the questionnaire had “not applicable” as an answer option. To be able to use the data from a run that had a “not applicable” response, all “not applicable” responses were entered as the affirmative of that answer group in the data spreadsheet. For example in the question “Is all necessary S&TE identified in the documentation?” the possible responses were: Correct, Incorrect and Not Applicable with all Not Applicable responses changed to Correct.

To enable the data to be coded for screening, each of the answers available in a question’s answer group were entered into a drop-down list. Using the list prevents typographical errors when entering data. Within the list, the available responses were entered from affirmative to negative responses. Maintaining the integrity of this process allowed for formulæ to be used in Microsoft® Excel to assign a number value to each possible response, with the affirmative responses assigned higher numbers.

Where the assessor’s comments provided an answer that contradicted the assessor’s selected response to circle, the data entered into the spreadsheet was based upon the comment. This decision was made for the reason that people get survey fatigue and may select the incorrect answer some of the time, however to write a response takes a conscious effort and thus the written word was given more weighting.
One important point that was identified was that only planned maintenance routines that were able to be completed, had assessments with all of the answers provided, that is a full data set. Where a routine contained incomplete responses, no data was recorded and the run deemed invalid.

When the historical data was first analysed in 2015-2016, without DOE techniques, each planned maintenance routine was assessed as suitable, suitable with deficiencies, or not suitable using summative data. The process of making the original summative assessments of each of the routines was not documented, but rather was based on assessors’ experience. A more consistent method of summative assessment was required to analyse the historical data. A Cooper-Harper rating scale (Cooper & Harper, 1969) was selected as a suitable scale to adapt. The modifications to the ratings scale were to focus on being able to assess a planned maintenance routine. The scale includes the ability of representative maintainers to be able to safely conduct the required maintenance according to the design intent, in the designated period of time, using the designated resources. The new assessment scale ranges from 0 to 9 as shown in two parts in Figures 4 and 5. An assessment of 0 meaning ‘not able to complete the maintenance’ and 9 meaning ‘the maintenance achieved exactly in accordance with the routine without any additional resources’. Resources include administration time, maintenance time, finances, materials, stores and external support (contractor and/or military subject matter experts). The extent of additional resources was codified as follows:

- Minimal – less than ten percent of additional resources.
- Some – more than ten percent additional resources but not more than 25 percent.
- Considerable – more than 25 percent additional resources but not more than 50 percent.
- Extreme – more than 50 percent additional resources.

The experience of the maintainers was considered in the scale. Comments within the data advised that the personnel experienced in maintaining equipment in their chosen profession were able to compensate for deficiencies in documentation and training. However, the routines should be written for the skill set expected of a person with the rank and work center as listed in the routine.
Safety factors were applied in two methods in the scale. The most severe assesses if the routine is conducted in accordance with the approved documentation whether it is likely to cause death; this has an automatic score of one (see Figure 5, the Part B). If there is a safety issue that is not likely to cause death then the rating is reduced by three. The adjusted rating is not a reflection of how serious the injury may be, just that there was a safety issue.

To ensure alignment of the scale with RANTEAA’s major recommendation categories in conducting suitability assessments for introduction to service, the numerical ratings within the scale were grouped
into the following summative assessment group:

- **Suitable – score 7-9** - remaining issues, if any, are minor and do not require resolution (i.e., issues are tolerable)
- **Suitable with deficiencies – score 5-6** - issues present cause increased workload to personnel or increased use of resources but can be managed for a period of time (i.e., issues should be resolved)
- **Not suitable – score 0-4** - issues present an unreasonable additional use of time or resources, an increased safety risk or the routine is not achievable (i.e., issues must be corrected)

As the process of making the original assessment was not known in detail, the assessment outcome using the new scale and process were compared carefully with:

- the historical data for each routine, including the individual questionnaire comments,
- the previous assessor’s comments on issues, and
- the historical final assessment.

**DATA CHECK**

Checks on the data for potential outliers found residuals were all less than the recommended 2.5 for the Studentized and less than the recommended two for Cooks-D (Reagan & Kiemele, 2008). Nonetheless, several high residuals were investigated with the majority of these being due to a run (PM routine) with a result at the extremes of the assessment scale. A number of high residuals were identified as data entry errors due to the incorrect selection in a drop-down list in the activities sheet, and this led to some data correction.

Data checks also found the following questions had insufficient variance, or were not contributing, and so these were removed from the analysis:

- Question 1.02 *Subjective Assessment of this Documentation?*
- Question 3.03 *Is S&TE suitable?*
- Question 5.03 *Are waste disposal solutions suitable for 90 days of operations?*

A correlation matrix was used to check for high correlation between the factors that might lead to multi-collinearity in the regression model (Reagan & Kiemele, 2008). The results showed that Question 6.02 *Are materials and stores codified?* and Question 6.03 *Do stores holdings support 90 days of maintenance?* had a correlation of 0.584. To remove issues of collinearity, the factor supporting Question 6.03 was removed from the model. Hence, 25 questions remained as factors for the final analysis.

**INTERACTIONS**

Initial regression modelling, without interactions, produced a very weak model with a very low regression ($R^2$) value as shown in Figure 6. Interactions or higher order relationships can improve weak models (Reagan & Kiemele, 2008), however, testing all of the interactions was not possible as it would require 2613 runs of valid data and there were only 279 runs of data.
Subject matter experts (SME) with experience in planned maintenance assessments were able to identify a total of 55 likely interactions plus ten quadratics of the three-level factors. The interactions are indicated by “X” in Figure 7.

**Figure 7:** SME identified interactions indicated by "X", with interactions in final model highlighted blue

There were 122 runs of data in the initial test period so the 65 interactions and quadratics were within the degrees of freedom of a multifactor regression. Nine of the chosen interactions and two of the chosen quadratics caused the model to fail due to insufficient data of particular combinations and so these were removed. All valid factors are highlighted blue in Figure 7. With the chosen and valid interactions included, the final model (Figure 8) was achieved by progressive removal of the likely least significant factors time and re-run of the multi-factor regression a total of 59 times (Schmidt & Launsby, 2005).
The inclusion of the interactions improved the model so that it explains about 47% of the variation in the data ($R^2$). However, the inclusion of the interactions also reduced many of the factors’ tolerances, with only one above 0.7 and only six above 0.5 (ideal of 1.0), such that the model is less orthogonal than the model without interactions. The standard error of 1.57, in a scale of 9, represents a high level of noise in the process. The limitations of the historical data make further understanding of the cause of the variation impossible without new testing.

The Pareto of coefficients (Figure 9) shows the factors and interactions that have the greatest effect on the suitability assessment.
While factors O, A and F have the greatest effect in their own right, when taking into consideration the significant interactions, such as OP and the five significant interactions involving factor G (i.e., GS, GP, GI, GU & GY), then factors O and G with their interactions account for 28 percent each of the overall effect in the supportability ratings, or 57% overall. This is shown in the surface plot at Figure 10, while the really important interactions of OP, GS and GP are shown in Figures 11 to 13. Just the four factors of O, G, A and F account for 75 percent of the overall effect, representing a significant ability to focus activities to prepare for supportability assessments when compared to the initial list of 29 factors.

The model could still be improved substantially, most likely with two-way interactions not tried or three-way interactions, all of which might be possible with increased test runs.

Figure 9: Pareto of Coefficients and Interactions (red – significant p<0.05, blue - likely significant 0.1<p<0.05, grey - non-significant p>0.1)

Figure 10: Surface plot of the two most significant factors (others held at mid-values of ranges)
VALIDATION

The validation of the model was conducted using the second set of historical data with 157 assessment
records. A histogram of the model’s residuals (actual value – predicted value) is shown in Figure 14. The standard deviation of just over two, is higher than the model’s predicted from the standard error of 1.57 earlier, and as an illustration, 44 percent of the 157 predictions lie within plus or minus one standard deviation of the actual score, which is less than a normal distribution would predict at 68 percent.

The standard deviation of the residuals was calculated for each assessment score to see if the model predicts better for good or poor assessments. Assessments of four or 7-9 all had residuals within the predicted standard deviation of 1.57, however, assessments of two and five had poorer predictive reliability with standard deviations of the residuals of greater than three. The validation confirms the model has predictive value but could be improved with more test runs and importantly, a better understanding through a measurement system analysis and DOE model with repetitions in order to understand the factors causing variation.

DISCUSSION OF SIGNIFICANCE

Pareto charts of the regression coefficients found that the following factors and their interactions, in order of importance, account for 75 percent of all effect on the overall supportability assessment:

- correct qualifications of maintainers to do the tasks (O),
- logic flow of maintenance instructions (G),
- documentation complete in order to finish activity (A), and
- accurately scheduling estimate for the task (F).

All these key factors and their important interactions can be verified and validated during ship developmental and acceptance testing using maintainability demonstrations and the metric of mean-time-to-repair. Such maintainability demonstrations were first proposed by Blanchard in 1974 and the U.S. DoD military standard 1388 (Australian version DEF AUST 5692) and repeated in Blanchard’s authoritative logistics texts ever since (Blanchard, 2013), as well as in teaching of logistics support analysis. The key to such demonstrations is to have them contracted from the outset and involving representative Navy maintenance personnel, which in turn critically depends on a front-end logistics support analysis to inform the tendering and contract terms.

FUTURE TESTING AND RESEARCH

RANTEAA is soon to conduct OT&E on SEA 4000-Air Warfare Destroyer; this will include suitability testing of the organic planned maintenance routines. The questionnaire used for these
assessments should be revised using lessons learned from conduct of this research. A measurement system analysis should be performed to determine the contribution of the measurement system to variance and from this training provided on the assessment method to personnel who will be interviewers so as to improve assessment consistency.

The questionnaire used for assessments should be updated to:

- remove correlated questions
- ask questions that are clearer for the interviewer and interviewee to reduce ambiguity in the answers and increase the orthogonality of the model
- add elements of actual administration and logistics time, and actual maintenance time to provide better defined experimental inputs, and
- include the assessment process flowchart developed in this research.

Future test design needs to have repetition built in, so that the source of variation can be better understood. Potentially, the simplification of the questionnaires to the more important factors to the overall assessment score (i.e., factors O, G, A & F), could gain efficiencies to enable planned maintenance routines to run twice; that is the minimum to get an understanding of variation. This research has contributed many valuable processes for inclusion in future data collection plans and standard processes for ship supportability T&E.

**CONCLUSION**

Research was conducted on historical data from supportability test and evaluation of Navy ships during their operational acceptance into service. The research used Design of Experiments test analysis techniques to identify which logistic and operational input factors have the greatest effect on the ability to complete planned maintenance of Navy ships. The 279 planned maintenance assessments used for the analysis and its validation were from two major RAN ships as part of three major introductory service milestones. There were 25 supportability factors assessed across each of the supportability assessments used to do the screening. A nine-point Cooper-Harper rating scale was adapted to code the output response. The data was analysed using multi-factor, multiple-response regression and due to the limited degrees of freedom the regression included only a subset of all possible two-factor interactions and quadratics as selected by logistics support specialists.

Nine supportability factors and eight two-factor interactions were found to be significant, represents a substantial focusing from all of the combinations possible in the current testing. However the current model does not have the accuracy required to estimate if the organic planned maintenance for a capability is, or is not, suitable based on input factor assessments alone. Pareto charts of the regression coefficients found that the factor of ‘qualifications of maintainers to do the maintenance tasks’ plus this factor’s interaction with maintainers’ ‘confidence in ability’ accounted for 28 percent of overall effect in the supportability ratings. Similarly the factor of ‘logic flow of maintenance instructions’ and this factor’s interactions with ‘realistic timing,’ ‘accessibility and ergonomics’, ‘confidence in ability’, ‘appropriate isolations’ and ‘spares availability and coding’, accounted for a further 28 percent of overall effect in supportability ratings. Both these key factors and their important interactions can be verified and validated during ship developmental and acceptance testing using maintainability demonstrations. Such maintainability demonstrations were first proposed by Blanchard and military standards in 1973-74 and repeated in authoritative logistics texts and logistics support training ever since. The key to such demonstrations is to have them contracted from the outset and involving representative Navy maintenance personnel.

The Royal Australian Navy is soon to conduct supportability assessments on the Hobart Class ships. The supportability assessment questionnaire is likely to be revised using lessons learned from this research analysis and checked for resolution and repeatability using a formal measurement system analysis, as taught in the course by the University of New South Wales.

**REFERENCES**

Ahner, D. K., 2016. ‘Better buying power, developmental testing, and scientific test and analysis
techniques,’ ITEA Journal, 37, pp. 286-290.
Defense Acquisition University (DAU), 2015. DAU T&E Workforce Development Process. DAU presentation by Professor Frank Kenlon, Mr Ed Verchot & Dr Paul Alfieri to Group Captain (Dr) K. Joiner of Australian Defence T&E Office and University of NSW, 17 August 2015.
RANTEAA JP 2048 Phase 4A IOC Suitability Assessment Questionnaires
RANTEAA JP 2048 Phase 4B OC1 Suitability Assessment Questionnaires

BIOGRAPHY
Lieutenant Commander Matthew Wernas joined the Royal Australian Navy in 1994 and spent nine years as an electronics technician submariner prior to becoming an electrical engineer. As an officer he has served in the Navy’s Adelaide class frigates and provided support to them during shore postings. He is currently the Integrated Logistic Support and Suitability Test Manager at the Royal Australian Navy Test, Evaluation and Acceptance Authority.

Dr Keith Joiner, CSC, joined the Air Force in 1985 and became an aeronautical engineer, project manager and teacher over a 30-year career before joining the University of New South Wales in 2015 as a senior lecturer in test and evaluation. From 2010 to 2014 he was the Director-General of Test and Evaluation for the Australian Defence Force, where he was awarded a Conspicuous Service Cross. He is a Certified Practising Engineer and a Certified Practising Project Director.