An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

by Samuel Velázquez

B.S. in Mechanical Engineering, May 1998, Univ. of Puerto Rico, Mayagüez Campus
M.S. in Mechanical Engineering, December 2004, Rensselaer Polytechnic Institute

A Praxis submitted to

The Faculty of
The School of Engineering and Applied Science
of The George Washington University
in partial satisfaction of the requirements
for the degree of Doctor of Engineering

May 19th, 2019

Praxis directed by

Muhammad F. Islam
Professorial Lecturer of Engineering Management and Systems Engineering

John M. Fossaceca
Professorial Lecturer of Engineering Management and Systems Engineering

The School of Engineering and Applied Science of The George Washington University certifies that Samuel Velázquez has passed the Final Examination for the degree of Doctor of Engineering as of March 15th, 2019. This is the final and approved form of the praxis.

An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

Samuel Velázquez

Praxis Research Committee:

Muhammad Islam, Professorial Lecturer of Engineering Management and Systems Engineering, Praxis Co-Director

John Fossaceca, Professorial Lecturer of Engineering Management and Systems Engineering, Praxis Co-Director

Amirhossein Etemadi, Assistant Professor of Engineering and Applied Science, Committee Member

Ebrahim Malalla, Visiting Associate Professor of Engineering and Applied Science, Committee Member

© Copyright 2019 by Samuel Velázquez
All rights reserved

Dedications

The author wishes to dedicate this praxis to his wife (his better half), Alma, and son, Gabriel. Their love, patience, and support were a constant drive in the pursuit of this degree, but mainly they are a source of inspiration and motivation that continues to guide me through hardship and life trials.

Acknowledgements

The author wishes to acknowledge the tremendous support and guidance provided by his advisors Dr. Muhammad Islam, Dr. John Fossaceca, Mr. Charles Bogue, and Dr. Amirhossein Etemadi. Additionally, the author would like to thank all the volunteers that donated their time and expertise to this research study.

Abstract

An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

Despite having risk management instructions and an official guidebook, risk assessments and analysis still lack the sufficient technical rigor when applying these processes to research and development projects within the Department of Defense (DoD). The risk analysis phase often suffers from expert judgment bias and siloed risk management decision-making approaches that can lead to projects not reaching their technical and programmatic goals. As a result, many of these processes are not implemented adequately, which impacts the product quality, mission, and customers of DoD engineering programs. This research attempts to supplement current DoD methodology with structured expert judgment (SEJ) and the use of a Bayesian inference methodology, by introducing a basic risk model into the DoD risk management workflow, to influence the use of scientific methods in the risk analysis phase. As a result, a typical application of the structured expert judgment method can outperform the current scoring method and have no negative impact on key program measures. A similar result is achieved when using the Bayesian inference is used to find the optimal risk probability distribution based on the observed data. Based on these findings it seems that the implication in practice is that the wrong methodology in the current DoD risk assessment is being used. Future work recommends the implementation of an organization level SEJ panel and expands efforts to make these improvements permanents into the current guidance. Further efforts are recommended to help with the data collection of observable data that can help the Bayesian inference analysis. As well to expand the research for additional optimal

Bayesian inference probability distributions capable of outperforming the current scoring method.

Table of Contents

Dedications ..... **iv**

Acknowledgements ..... **v**

Abstract ..... **vi**

List of Figures ..... **xi**

List of Tables ..... **xiii**

List of Equations ..... **xvi**

List of Acronyms ..... **xviii**

Chapter 1 Introduction **1**

1.1 Background 1
1.2 Statement of the Research Problem 3
1.2.1 Research Problem Elaboration 3
1.3 Thesis Statement 3
1.3.1 Thesis Elaboration 4
1.4 Research Questions 5
1.5 Research Hypotheses 5
1.6 Objective, Assumptions, and Limitations 8
1.7 Expected Findings 10

Chapter 2 Literature Review **11**

2.1 A Critical Scoring Method Review 11
2.2 DoD Guidelines for Risk Analysis 19
2.3 Engineering Management Body of Knowledge 24
2.4 Overview of Cooke's Classical Method 26
2.4.1 Assumptions 30
2.4.2 Limitations 30
2.5 An Overview of the Bayesian Inference Method 31
2.5.1 Assumptions 33
2.5.2 Limitations 34

Chapter 3 Research Methodology **35**

3.1 Decision-Making Process Overview 35
3.2 Tailored Elicitation (Based on Classical Method) 37
3.2.1 Pre-Calibration 37
3.2.1.1 Obtaining Project Data 37
3.2.1.2 Data Review and Categorization 38

3.2.1.3 Experts Selection 40
3.2.1.4 Initial Survey and Consent 40
3.2.2 Calibration Phase 41
3.2.2.1 Expert's Training 41
3.2.2.2 Dry-Run 48
3.2.2.3 Calibration Survey 48
3.2.3 Elicitation Phase 50
3.2.4 Post-Elicitation 51
3.2.4.1 Discrepancy Test 51
3.2.4.2 Robustness Test 52
3.3 Data Analysis Method (Evidence-Based) 52
3.3.1 Bayesian Inference (BI) 54
3.3.1.1 Binomial and Beta Distributions 55
3.3.1.2 Poisson and Gamma Distributions 56
3.3.1.3 Exponential and Gamma Distributions 57
3.3.1.4 Weibull and Lognormal Distributions 58
3.3.1.5 Model Checking with Summary Statistics and Bayesian Information 58
3.3.1.6 OpenBUGS Modeling Scripts 60
3.3.1.7 Hardware Codes 60
Chapter 4 Results 61
4.1 Expert Assessments and Data Collection 61
4.2 Expert Performance 67
4.2.1 Expert's Robustness Results 76
4.2.2 Expert's Discrepancy Results 77
4.3 Bayesian Inference Results 79
4.4 Hypothesis Testing 82
4.4.1 Testing H1 and H2 (Program Metrics) 82
4.4.2 Testing H3 (SEJ) 87
4.4.3 Testing H4 (BI) 88
Chapter 5 Conclusions and Recommendations 91
5.1 Discussion on Research Question RQ1 91
5.2 Discussion on Research Question RQ2 92
5.3 Discussion on Research Question RQ3 93
5.4 Summary of Hypotheses Outcomes 95
5.5 Recommendations (Implications) For Practice 95

5.6 Future Work 96
References 99
Appendix A: Calibration Survey 103
Appendix B: Elicitation Survey 116
Appendix C: Consent Form 125
Appendix D: Consent Form 128
Appendix E: WinBUGS Scripts 135
Appendix F: Range Graphs 138

List of Figures

Figure 1 - Example of Risk Matrix with Scoring 2

Figure 2 - Proposed Changes to Basic Risk Model due to Praxis Research 9

Figure 3 - A Standard Risk Matrix as used by the FAA (Cox, 2008) 14

Figure 4 - DoD Acquisition Life Cycle (DASDSE, 2017) 19

Figure 5 - DoD Risk Analysis Phase (DASDSE, 2017) 20

Figure 6 - Typical Risk Reporting from DoD RIO guidebook (DASDSE, 2017) 23

Figure 7 - EMBOK Probability/Impact Matrix 25

Figure 8 - PRA Cost Decision Matrix (Lund, 2008) 33

Figure 9 - Basic Risk Model (Expanded) 36

Figure 10 – Body of Evidence Approach (Kaplan, 1997) 37

Figure 11 - Risk Exposure Tool (Response Scale) 49

Figure 12 - Expert Assessment in Excalibur 51

Figure 13 - Option GUI during Excalibur Run 52

Figure 14 - Flow diagram for analysis method (Kelly and et al. 2009) 53

Figure 15 - Example of a Well-Calibrated Expert 68

Figure 16 - Run parameters used for Excalibur 68

Figure 17 - Calibration Results on Adapter Quality 70

Figure 18 - Calibration Results on Funding Award 71

Figure 19 - Calibration Results on Draft RMF Requirement 73

Figure 20 - Calibration Results on Placing Raw Materials Orders 73

Figure 21 - Elicitation Results on Widget Malfunction 74

Figure 22 - Elicitation Results on Reliability 74

Figure 23 - Elicitation Results on Communication..... 75
Figure 24 - Elicitation Results on Technical 75
Figure 25 - Discrepancy Analysis Chart 79
Figure 26 - Normality Plot for Research Cost 84
Figure 27 - Normality Test for Research Time..... 85
Figure 28- Ideal Framework for Basic Risk Model with Organization Level Support 97
Figure A - 1 – Adapter Quality Over Time 105
Figure A - 2 – Pixilation Levels..... 107
Figure A - 3 – CPU Power..... 108
Figure A - 4 – Pixilation Level Over Time 108
Figure B - 1 – Widget Temperature 118
Figure B - 2 – Electrical Degradation of Older HDD Model 119
Figure B - 3 – TPM Chart..... 123
Figure E - 1 – Exponential OpenBUGS Script..... 135
Figure E - 2 – Weibull OpenBUGS Script 136
Figure E - 3 – Lognormal OpenBUGS Script 137

List of Tables

Table 1 - Hypotheses Summary 7

Table 2 - Weight Score Example 13

Table 3 - A Standard Risk Matrix as used by the FAA (Cox, 2008) 14

Table 4 - Source NIST800-30 p21 - Likelihood Level (Ross, 2012) 16

Table 5 – Source NIST300-30 p23 - Impact Levels (Ross, 2012) 16

Table 6 - Typical Consequence Table from DoD RIO guidebook (DASDSE, 2017) 21

Table 7 - Typical Likelihood Table from DoD RIO guidebook (DASDSE, 2017) 22

Table 8 - Risk Prioritization Example 23

Table 9 - SEJ Raw Data Sources (Summary) 38

Table 10 - Calibration Phase Seed Variables with Realizations 38

Table 11 - Elicitation Phase Target Variables (*Variables of Interest*) 39

Table 12 - Research Study Experts (Spartan Team) 40

Table 13 - Spartan Team w/ Pay Grade Classification 41

Table 14 - Bayes Theorem Elements 54

Table 15 - Hardware Codes from Project P4 60

Table 16 - Complete SEJ Assessment for Experts S117 to S132 62

Table 17 - Summary of Experts Performance (Seed Variables) 69

Table 18 - Summary of Experts Performance (Target Variables) 71

Table 19 – Expert Robustness Results 76

Table 20 – Seed Variables Robustness Results 77

Table 21 - Discrepancy Analysis Results 78

Table 22 - Exponential Model Summary Statistics for Widgets A thru D 80

Table 23 - Weibull Model Summary Statistics for Widgets A thru D 80
Table 24 - Lognormal Model Summary Statistics for Widgets A thru D 81
Table 25 - Bayes Factor Results on Summary Statistics 81
Table 26 - Bayesian Information Criterion (BIC) Results 82
Table 27 - Cost and Time Research Data Collected 83
Table 28 – Mean Baseline Cost from P1, P2, and P3 86
Table 29 – Mean Baseline Time from P1, P2, and P3 86
Table 30 - Student T-Test Results for Research Time 86
Table 31 - Student T-Test Results for Research Cost 87
Table 32 - Summary of PWDM Performance 88
Table 33 - Bayes Factor Acceptance Criteria for Hypothesis Test 89
Table 34 - BF Acceptance Criteria 89
Table 35 - Hypothesis Testing Results 95
Table A - 1 – Risk Exposure Tool (Calibration Survey) 103
Table A - 2 – Calibration Answer Item #1 104
Table A - 3 – Calibration Answer Item #2 106
Table A - 4 – Calibration Answer Item #3 107
Table A - 5 – Calibration Answer Item #4 109
Table A - 6 – Calibration Answer Item #5 110
Table A - 7 – Calibration Answer Item #6 111
Table A - 8 – Calibration Answer Item #7 112
Table A - 9 – Calibration Answer Item #8 113
Table A - 10 – Calibration Answer Item #9 114

Table A - 11 – Calibration Answer Item #10 115

Table B - 1 – Risk Exposure Tool (Elicitation Survey) 116

Table B - 2 – Elicitation Answer Item #1..... 118

Table B - 3 – Elicitation Answer Item #2..... 120

Table B - 4 – Elicitation Answer Item #3..... 121

Table B - 5 – Elicitation Answer Item #4..... 124

List of Equations

Equation 1 - Expert's Inter-Quartile Intervals 44
Equation 2 - Seed variables frequency 45
Equation 3 - Probability distribution 46
Equation 4 - Sample Distribution Equation 46
Equation 5 - Calibration Score Equation 46
Equation 6 - Information Score Equation 47
Equation 7 - Global Weight Equation 47
Equation 8 - Combined expert scores (DM) 47
Equation 9 - Bayes Theorem 54
Equation 10 - Binomial Distribution 55
Equation 11 - Beta Distribution 55
Equation 12 - Jeffrey's Prior Parameters for Beta Distribution 56
Equation 13 - Poisson Distribution 56
Equation 14 - Gamma Distribution 56
Equation 15 - Jeffrey's Parameters for a Gamma Distribution 57
Equation 16 - Exponential Distribution 57
Equation 17 - Jeffrey's Parameters for a Gamma Distribution (Exponential Case) 57
Equation 18 - Weibull Distribution PDF 58
Equation 19 - Lognormal Distribution 58
Equation 20 - Chi-square test statistic for observed values 58
Equation 21 - Chi-square test statistic for predicted values 59
Equation 22 - Bayes Factor Definition 59

Equation 23 - Bayesian Information Criterion (BIC) 60

List of Acronyms

APB	Acquisition Program Baseline
APUC	Average Procurement Unit Cost
AZ	Grade Code Assigned to Experts
BI	Bayesian Inference
BZ	Grade Code Assigned to Experts
CDRL	Contract Data Requirements List
CMP	Configuration Management Plan
CONOPS	Concept of Operations
CP	Calibration Power
CPU	Central Processing Unit
CTP	Critical Technical Parameter(s)
DM	Decision Maker
DoD	Department of Defense
EMBOK	Engineering Management Book of Knowledge
EMSE	Engineering Management and Systems Engineering
EW	Equal Weight
FAA	Federal Aviation Administration
FHA	Federal Highway Administration
FTE	Full Time Equivalent
FY	Fiscal Year
GPU	Graphical Processing Unit
H	High Probability

H_0	Null Hypothesis
H_a	Alternate Hypothesis
ICD	Interface Control Document
IRS	Interface Requirements Specification
KPP	Key Performance Parameter
KSA	Key System Attribute
L	Low Probability
M	Medium Probability
MCMC	Markov Chain Monte Carlo
MP	Major Program
NASA	National Aeronautical Space Administration
NIST	National Institute of Standards and Technology
NX	Grade Code Assigned to Experts
OMS	Operations and Maintenance Services
OpenBUGS	Bayesian inference Using Gibbs Sampling (Open source)
P1	Workplace Project #1 - Actual Project Name Removed
P2	Workplace Project #2 - Actual Project Name Removed
P3	Workplace Project #3 - Actual Project Name Removed
P4	Workplace Project #4 - Actual Project Name Removed
PAUC	Program Acquisition Unit Cost
PEO	Program Executive Office
PMBOK	Project Management Book of Knowledge
PMI	Project Management Institute

PoR	Program of Record
PRA	Probability Risk Assessment
PW	Performance Weight
QRA	Quantitative Risk Assessment
R**&**D	Research and Development
RDT&E	Research, Development, Test & Evaluation
RE	Risk Exposure
RIO	Risks, Issues, and Opportunities
RM	Risk Management
RMF	Risk Management Framework
RMP	Risk Management Plan
S117 to S132	Expert Codes given to volunteers to maintain research study participation anonymous
SEJ	Structured Expert Judgment
SETR	Systems Engineering Technical Review
SPSS	Statistical Package for the Social Sciences
SSS	System Requirements Specification, System/Subsystem Specification
TPM	Technical Performance Measure
VH	Very High Probability
VL	Very Low Probability
WCE	Well-Calibrated Expert

Chapter 1 Introduction

Risk management (RM) consists of the identification of risks, its assessment, and analysis, the effective planning and execution of mitigation strategies to reduce the assessed impact to closure to minimize the recurrence of the identified root cause. Within the context of RM and this praxis, a risk is an uncertain event defined by its probability and impact to the project or program. The risk analysis phase is characterized by providing quantitative or qualitative output and establishing the link between root cause and closure criteria for the mitigation strategy.

1.1 Background

Risk analysis is the cornerstone of any risk management strategy. In his book, *The Failure of Risk Management*, Hubbard is relentless in his message when he suggests that the most prominent failure of risk management is that most processes do not measure or validate the methods that are required to incorporate empirical science into risk management (Hubbard D. W., 2009). Aspects of risk analysis phases which include the risk matrix (risk cube) are not rooted as profoundly as the scientific method and empirical evidence but are typically considered the cornerstone of the analysis. A typical risk matrix and scoring method is depicted in Figure 1. *This figure is used only for reference.*

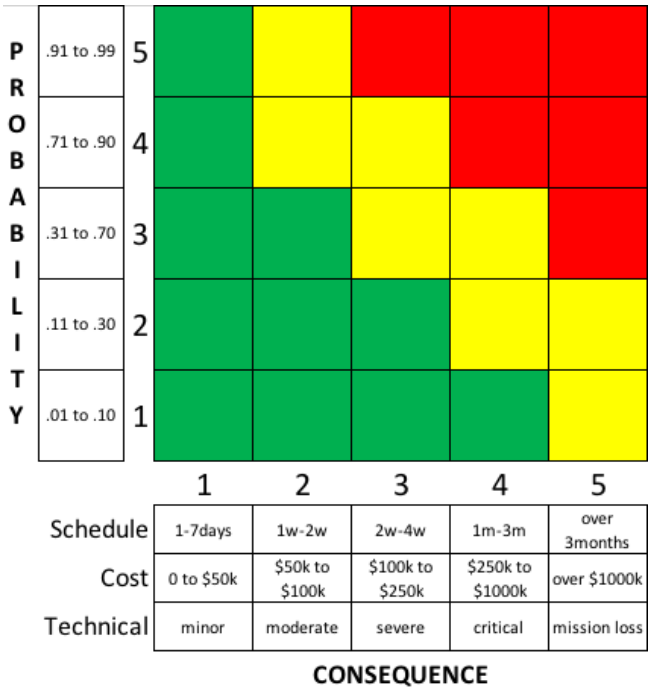


Figure 1 - Example of Risk Matrix with Scoring

Cox suggests that the risk matrix was widely adopted for its simplicity; however, he made a distinction that the "*...knowledge gap often exhibited in these matrices implies that the actual performance of the risk matrix and whether it is helpful, is no better than random*" (Cox, 2008, p. 500). In this context, the use of a risk matrix, similar scoring methods and any lack of empirical control or evidence when performing a risk analysis seems futile (Cox, 2008). This research study focus is on improving the risk analysis phase within the current DoD risk management process. This research discusses a method to enhance the subjective part of the basic risk model used during the risk analysis. The proposed method accounts for the uncertainty and subjectivity based on the lack of data in the risk analysis using the Structured Expert Judgment (SEJ) method. A second method which characterizes the quantitative part of the basic risk model. This method, based on Bayesian Inference (BI), and uses available data to the study to construct and compare

probability distributions based on the data type to determine the most representative to the data at hand.

1.2 Statement of the Research Problem

The problem statement for this research study is the following:

- *The current scoring method used within the DoD risk management workflow lacks the methodology within its basic risk model to evaluate an objective or subjective risk analysis.*

1.2.1 Research Problem Elaboration

Current Department of Defense (DoD) policy instruction 5000.02 and document Risk, Issue and Opportunity (RIO) guidebook provides guidance on how to conduct a risk analysis at a fundamental level but fails to describe what risk model and processes to use to perform an objective and subjective risk analysis. Consequently, the resulting risk analysis is inadequate to characterize the probabilistic uncertainty correctly, which can eventually impact the technical and programmatic goals of any DoD program. Either overstating or understating a risk analysis can lead to mitigation strategies with incorrect resources, the misaligned scope, and closure criteria that will not satisfy the original intent. The risk analysis phase often suffers from expert judgment bias and siloed risk management decision-making approaches that can lead to a decline in technical and programmatic goals. As a result, many of these processes are not implemented adequately, which impacts the product quality, mission, and customers of DoD engineering programs.

1.3 Thesis Statement

The thesis statement for this research study is the following:

- *Both the structured expert judgment (SEJ) and Bayesian inference (BI) methodologies provide the probability and modeling framework needed to build the basic risk model within the DoD risk management workflow.*

1.3.1 Thesis Elaboration

The DoD risk analysis phase often lacks sufficient technical rigor and empirical control to develop risk assessments that are not expert judgment biased. Likewise, the decision-making process often gets impacted by a siloed risk assessment. Both conditions have a direct contribution and effects to the technical and programmatic goals. The proposed solution will provide the practitioner the tools, procedures, and guidance to handle a risk model with subjective probabilities and uncertainty, and the data analysis capabilities to ensure adequate risk assessments without impacting project cost, schedule and performance metrics.

When assessing uncertainties during a risk assessment, risk matrices and scoring systems do not yield reliable results due, in part, to the constant bias exhibited when evaluating likelihoods and consequences as observed by (Smith, Siefert, & Drain, 2008). Both Cox and Hubbard have extensive research showing how these scoring methods fail to measure or achieve consensus on a risk assessment's uncertainty (Cox, 2008) (Hubbard D. W., 2009). So why rely on them? Why not use other methods to determine these values empirically and more accurately? That is the proposed solution and argument of this research study. When using the scientific method, the SEJ method, and the BI method to come to an agreement based on the empirical data, a set of assumption and hypotheses, the result is vastly improved to those resulting from risk matrices and scoring systems.

1.4 Research Questions

This research was undertaken to improve the risk analysis phase in the DoD risk management workflow. Aims to answer the following fundamental questions on how the improvement could affect the risk analysis and critical project metrics. For instance, the first and second questions focus on the statistical significance comparison between the status-quo and the proposed changes in the risk model:

- RQ1: Is there a cost or time impact when implementing the SEJ process within the DoD risk assessment workflow?
- RQ2: Will the statistical significance for the SEJ performance-weighted decision-maker (PWDM) outperform that of a traditional DoD risk assessment?

The last question explores how the proposed changes may impact the process and two key project metrics: cost and time:

- RQ3: Will the BI model predictions be enough to outperform that of a traditional DoD risk assessment, based on the available data provided?

1.5 Research Hypotheses

To correctly answer the research questions postulated in section 1.4 the following hypotheses are defined to support the fact-finding and empirical data collection. These hypotheses will serve as the building blocks for the research method and the data to be tested in Chapter 4.

- **Statement No. 1** suggests that by using the SEJ method within the context of the DoD risk management workflow will not have negative impacts on programmatic metrics typically tracked by projects, such as cost and time. Hence the postulation of the following hypotheses 1 and 2.

- **Hypothesis No. 1:**
    - $H1_0$: The average time to perform an SEJ risk analysis will be greater than or equal to the traditional risk analysis.
    - $H1_a$: The average time to perform an SEJ risk analysis will be less than the traditional risk analysis.
- **Hypothesis No. 2:**
    - $H2_0$: The average cost to perform an SEJ risk analysis will be greater than or equal to the traditional risk analysis.
    - $H2_a$: The average cost to perform an SEJ risk analysis will be less than the traditional risk analysis.
- **Statement No. 2** suggests that pooling of performance weighted experts based on seed variables or as the SEJ refers to PWDM, the statistical significance of the Cooke's classical model will be quantifiably better than that of the traditional scoring method using in the DoD risk management workflow. Hence the postulation of the following hypothesis 3.
- **Hypothesis No. 3:**
    - $H3_0$: The use of the SEJ method will not reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.
    - $H3_a$: The use of the SEJ method will reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.
- **Statement No. 3** suggests that when the BI method is used with probabilistic risk assessments with observed data, in this case from Project P4, it could predict values

of the probability distribution based on the available evidence or data available. Hence the postulation of the following hypothesis 4.

- **Hypothesis No. 4:**
    - $H4_0$: On average there is a difference between the predicted and the observed values when using the Bayes inference probability distribution model.
    - $H4_a$: On average there is no difference between the predicted and the observed values when using the Bayes inference probability distribution model.

Finally, all the hypotheses are summarized in Table 1.

Table 1 - Hypotheses Summary

Null Hypothesis (H_0)	Alternate Hypothesis (H_a)
The average time to perform an SEJ risk analysis will be greater than or equal to the traditional risk analysis.	The average time to perform an SEJ risk analysis will less than the traditional risk analysis.
The average cost to perform an SEJ risk analysis will be greater than or equal to the traditional risk analysis.	The average cost to perform an SEJ risk analysis will be less than the traditional risk analysis.
The use of the SEJ method will not reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.	The use of the SEJ method will reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.
On average there is a difference between the predicted and the observed values when using the Bayes inference probability distribution model.	On average there is no difference between the predicted and the observed values when using the Bayes inference probability distribution model.

1.6 Objective, Assumptions, and Limitations

Primarily, this praxis provides the methods required to improve the basic risk model underlining the current risk analysis phase within the larger DoD risk management process. This study leverages well-known and established principles to handle subjective probability and uncertainty with SEJ and the objective analysis using BI. Figure 2 illustrates the proposed changes outlined in this praxis. The primary research objective will focus on answering the research questions by collecting empirical data to reject or accept each hypothesis which is defined as part of the research methodology in Chapter 3. A fundamental assumption in this praxis is that both the BI and SEJ methods will help answer all the research questions; this praxis will also use evidence to assert SEJ's improvements for the DoD risk management process.

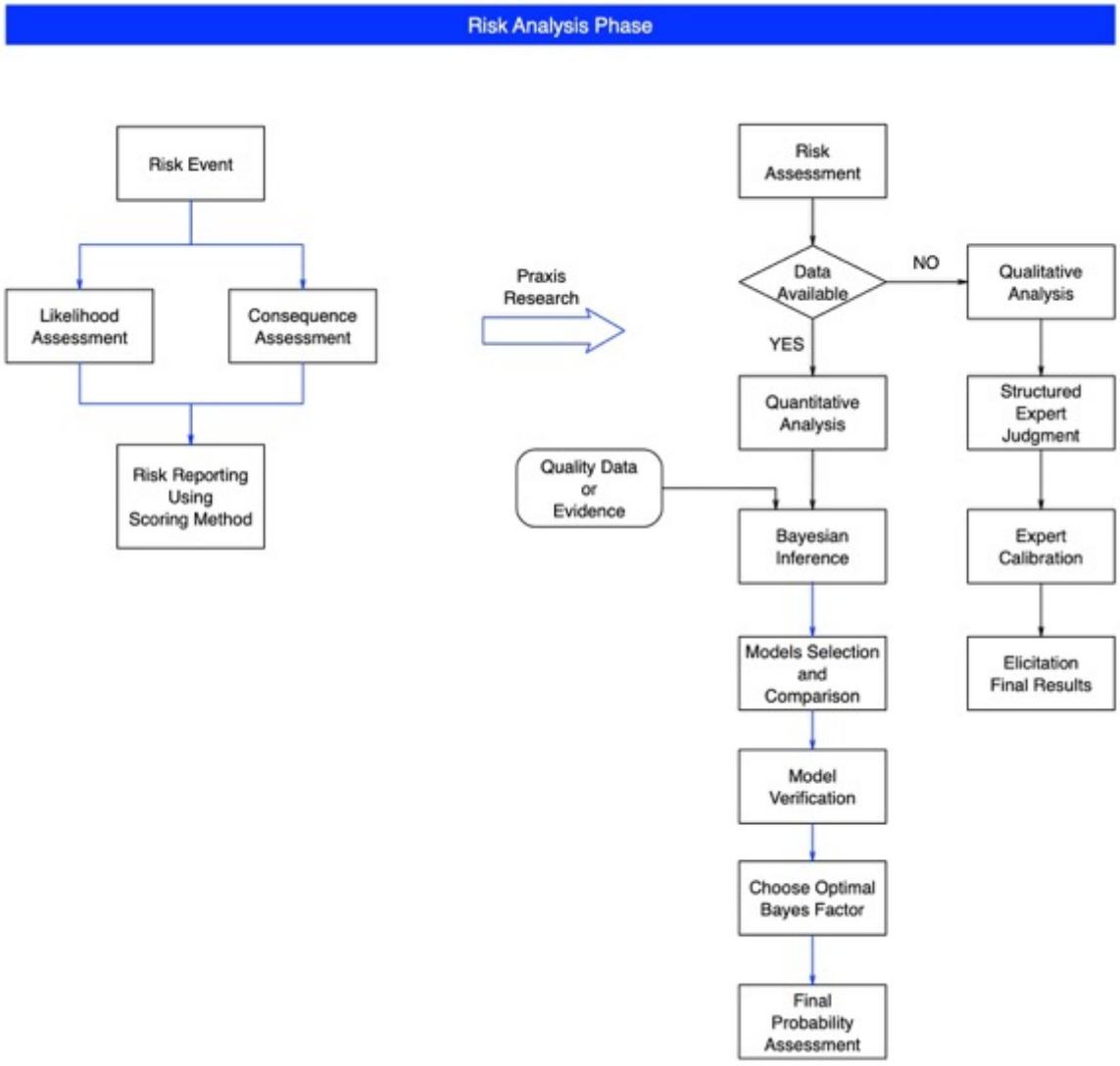


Figure 2 - Proposed Changes to Basic Risk Model due to Praxis Research

This research focuses on the process improvement to perform a risk analysis phase within the current DoD risk management process using the SEJ method with donated project data from Projects P1, P2, and P3. The following assumptions have been made:

- The volunteers supplying the subjective assessments will do so in their related field(s) of expertise.
- The observed data holds relevant information to predict the probability distribution.

- The donated data can be used to baseline the time and cost metrics.

Specific limitations of this research study will be discussed in detail in sections 2.4.2 and 2.5.2.

1.7 Expected Findings

The expected findings of this research study are as follows:

- The SEJ method will enable DoD risk management practitioners to improve their risk analysis while reducing the cost and duration required to complete the analysis.
- The SEJ method will enable DoD risk management practitioners to improve risk exposure and probability assessments based on the tools, metrics, and procedures used.
- Using the BI method and the observable data an optimal model probability distribution can be determined.

Chapter 2 Literature Review

This chapter provides an overview of the structured expert judgment method, and the Bayesian inference methods as well as those by the EMBOK and DoD RIO guidebook. It also reviews cognitive bias and heuristics in risks and the effects it has on the decision-making process and explores the assumptions and limitations in the research methodology.

2.1 A Critical Scoring Method Review

Scoring methods are widely used because of the convenience they provide. Unfortunately, their utility ends there as even the most elaborately constructed scoring method cannot seem to avoid one of the most fundamental problems, which is almost always developed in isolation or within this siloed effect during the risk assessment, in the risk analysis (Hubbard D. W., 2009). Risk analysis failure is not often understood by the practitioner and stakeholders; this is partly because the siloed result does not consider the data, information, or expertise of everyone involved in the process. Therefore, the scoring method cannot account for different perceptions of risks and uncertainties. According to Hubbard, risk analysis failure is attributed to three main reasons:

- The first reason is the failure to measure and validate data. Hubbard elaborates that it is essential for effective risk analysis that practitioners complete the experimental verification of data for the method used (Hubbard D. W., 2009).
- The second reason is the incorrect verification of data. Any siloed, subjective evidence or opinion made towards risk analysis success is believed to have no bearing in the success of the risk. There are instances in which experimental verification of data is completed by subjective means. Unfortunately, if care is not

taken, the data verification can lead to biased results due to unstructured and faulty errors in the methods used to vet the experts providing the data verification.

- The last reason is introducing unnecessary ambiguity because of ignoring data or methods by which data is collected. Hubbard explains that there are methods that work in real life situations just as well as controlled environments (Hubbard D. W., 2009).

To avoid the pitfalls often found in risk assessments due to the use of ordinal scales, Hubbard & Evans recommend the following criteria for proper risk assessments:

- Must use explicit, quantitatively-expressed probabilities and magnitudes for losses instead of surrogate nominal scales. One principal argument against the use of quantitative expressed probabilities and losses is the lack of data. Both Hubbard and Evans make a compelling counter-argument: Ordinal scales do not alleviate the problem of a lack of data; they merely mask it. With quantitative probabilities, however, the uncertainty that comes with a lack of data is explicitly incorporated in the estimate and refined over time (Hubbard & Evans, 2010).
- When using simulation techniques, such as Monte Carlo simulations, the model needs to correlate explicitly to the system or sub-component and the bounding event.

These scoring methods, take the approach of using subjective point scales to represent the problem. Among scoring methods, there is the weighted score variant, which, by design, uses a series of elaborate risk indicators to determine a weighted score of the investigated problem. In this instance, practitioners assign weighted scores from 0.1 to 1.0 to an event. The subjective nature of selecting a weighted value makes this method just as

reliable as the classical scoring method with no bearing on the final decision (Hubbard D. W., 2009). Table 2 shows a typical weighted scoring example.

Table 2 - Weight Score Example

Risk Event	Weight	Probability	Weighted Probability
On Time Delivery	0.90	0.74	0.67
Product Quality	0.50	0.55	0.28
Component Quality Improvements	0.75	0.25	0.19
Factory Cost Reductions	0.25	0.50	0.13
Design Changes	0.90	0.80	0.72

Almost all scoring methods are based on some ordinal scale - a scale made of arbitrary "units" with no relative order. Because of this, ordinal scales cannot be treated as a unit of measure, and hence no arithmetic can be performed (Hubbard D. W., 2009). Cox concludes that risk matrices, a key trait of many scoring methods, suffer from poor resolution, errors, suboptimal resource allocations, and ambiguous inputs and outputs (Cox, 2008). As Hubbard and Evans observed risk matrices come in various shapes and formats, and since the matrices are nominal scales their creation is arbitrary (Hubbard & Evans, 2010) . For example, The Federal Highway Administration (FHA) uses a risk matrix and scale for both the probability and impact from very low (VL) to very high (VH) as Table 3 shows.

Table 3 - A Standard Risk Matrix as used by the FAA (Cox, 2008)

Probability/Impact	VL	L	M	H	VH
VH	Green	Yellow	Red	Red	Red
H	Green	Yellow	Red	Red	Red
M	Green	Green	Yellow	Red	Red
L	Green	Green	Yellow	Red	Red
VL	Green	Green	Green	Yellow	Red

By comparison, The Federal Aviation Administration (FAA) uses a risk matrix with verbal labels for the probability and impact as Figure 3 demonstrates.

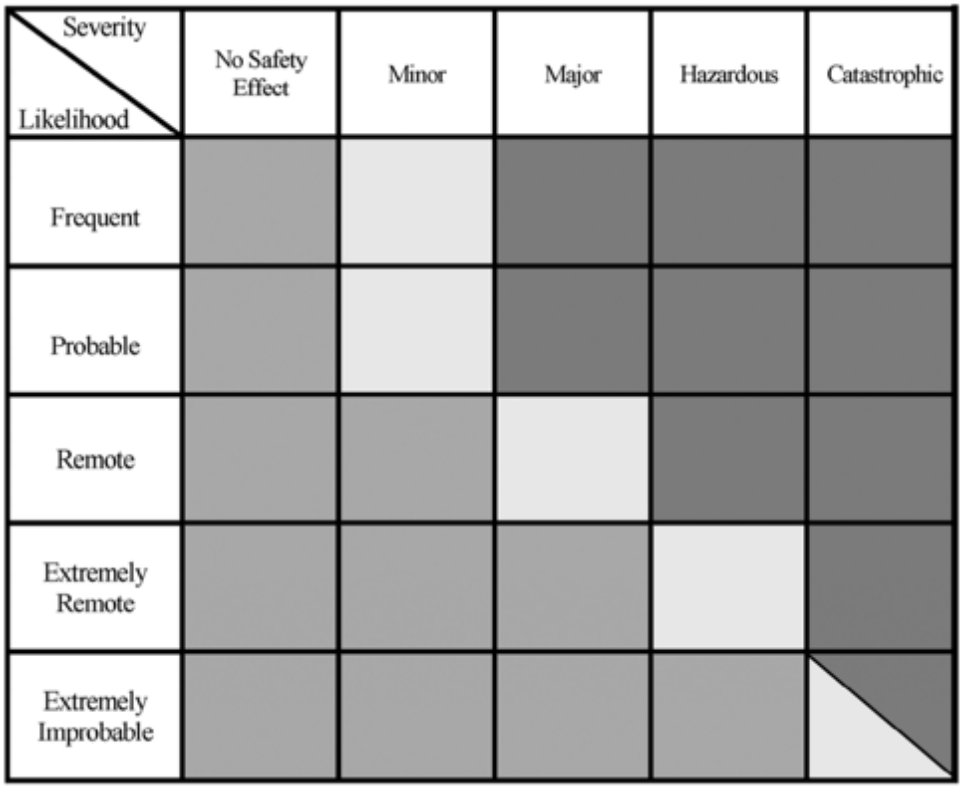


Figure 3 - A Standard Risk Matrix as used by the FAA (Cox, 2008)

Both examples, and the previous one depicted in Figure 1 support the findings made by (Cox, 2008). These examples also highlight the flaws in risk matrices when deviating from a set of logically consistent rules explained by (Thomas & Bratvold, 2014). Both

Thomas and Bratvold argue with the logical consistency examined in the field it failed to follow a consistent coloring assignment for "red" or "yellow" risks assigned to these categories (Thomas & Bratvold, 2014). Ball and Watt concluded that psychosocial factors, beliefs, and other views affect the context and may create an untrustworthy picture of the risks (Ball & Watt, 2013). Research seems to suggest that risk matrices cannot provide adequate resource allocations or estimates if they continue to be created in isolation with less than satisfactory and untrustworthy inputs. Nevertheless, projects and programs still rely on this tool exclusively for risk assessments and exposures, resource allocation, and critical planning that helps make these very same decisions on a daily basis without question.

When performing standard risk assessments with the current scoring methods, research shows that subjective risk assessments, in general, can be categorized into additive and multiplicative scores (Hubbard & Evans, 2010). Independent factors assessed during the analysis describe the additive score. Practitioners typically weight these scores and add them together. A multiplicative score is characterized by having two-to-three independent factors, which the practitioners subsequently multiply together. It is the multiplicative category that best describes the scoring method used by DoD. Interestingly, the National Institute of Standards and Technology (NIST) created standard NIST800-30 using a similar scoring system as the multiplicative score, and it is considered "best practice" (Hubbard & Evans, 2010). For comparison, Table 4 and Table 5 define the NIST800-30 standards for likelihood and consequence (impact), respectively.

Table 4 - Source NIST800-30 p21 - Likelihood Level (Ross, 2012)

Likelihood Level	Likelihood Definition
High	The threat-source is highly motivated and sufficiently capable, and controls to prevent the vulnerability from bringing exercised are ineffective.
Medium	The threat-source is motivated and sufficiently capable, but controls are in place that may impede the successful exercise of the vulnerability.
Low	The threat-source lacks motivation or capability, or controls are in place to prevent, or at least significant impede, the vulnerability from being exercised.

Table 5 – Source NIST300-30 p23 - Impact Levels (Ross, 2012)

Magnitude of Impact	Impact Definition
High	Exercise of the vulnerability (1) may result in the highly costly loss of major tangible assets or resources; (2) may significantly violate, harm, or impede an organization's mission, reputation, or interest; or (3) may result in human death or serious injury.
Medium	Exercise of the vulnerability (1) may result in the costly loss of tangible assets or resources; (2) may violate, harm, or impede an organization's mission, reputation, or interest; or (3) may result in human injury.
Low	Exercise of the vulnerability (1) may result in the loss of some tangible assets or resources or (2) may affect an organization's mission, reputation, or interest.

There are documented advantages of using scoring methods, among these are that scoring methods provide information relatively quickly, promote fast discussions, and provide an understandable graphical representation of risk exposure (Peace, 2017). As discussed previously, however, scoring methods do not yield reliable results when assessing uncertainties during risk assessments in part because the constant bias exhibited when evaluating likelihoods and consequences (Smith, Siefert, & Drain, 2008). According

to Hubbard and Evans, there are four main biases with scoring methods: cognitive bias, verbal labels, invalid inferences, and invisible correlations (Hubbard & Evans, 2010).

When evaluating risks, uncertainty, and probabilities, all practitioners are ineffective to some degree. Haselton, Nettle, and Andrews defined this behavior, or cognitive bias, as a systematic pattern deviation from rationality in judgment (Haselton, Nettle, & Andrews, 2005). In other words, there are significant implications to employ scoring methods with cognitive bias and to the risk assessment output. As Hubbard and Evans note, psychologists have found many other cognitive biases that can negatively influence a risk assessment (Hubbard & Evans, 2010). Baybutt defines heuristics as a mental shortcut taken by humans when applying the decision-making process (Baybutt, 2017). Examples of heuristics include educated guesses, the rule of thumb, and trial and error. As Baybutt explains, cognitive bias can surface from heuristics, but it can also arise from motivational biases, which are biases that favor a particular outcome (Baybutt, 2017). Some of the most common heuristics and cognitive biases to affect risk analysis include the ambiguity effect, anchoring, authority bias, groupthink, halo effect, and peer pressure (Baybutt, 2017). In general, research shows that cognitive biases influence the decision-making process negatively.

One characteristic all scoring methods share is the use of labels to determine probabilities or impact range, which practice elicit judgment. Unfortunately, when using these labels, the expected interpretation will be different for every elicitation depending on the practitioner's individual experiences, recollection, and cognitive bias. Scoring methods often come with additional features, scales, or schemes that are intended to aid in the risk assessment. The careless employment of these features leads to invalid inferences

(Hubbard & Evans, 2010). Similarly, when a risk assessment's resulting scores are combined with each other, invalid correlations could be deduced from the combination. Not only this combination may not follow probability laws, but the events may not relate. Hubbard and Evans found four scoring methodology fallacies:

- The first fallacy suggests that expert's subjective assessments are adequate assessments for risks. The first fallacy statement is proven false using cognitive bias behavior for any given risk assessment.
- The second fallacy shows that scoring methods scales, and labels have some "technical" rigor to their definition. The second fallacy statement is proven false as the labels, scales, and ranges are highly inconsistent and open to interpretation.
- The third fallacy suggests that scoring method's calculations are arbitrary and without any implication to the experts' elicitations. The third fallacy statement is proven false as the combination of these arbitrary scales can lead to invalid combinations and correlations.
- The fourth fallacy regards the concept that correlations between events, especially during risk analysis are not needed to complete a rough approximation. When performing a quantitative risk assessment (QRA), not accounting for correlations of related risks are not or not including failure data is in the modeling, will exceedingly underestimate the assessed risk. The last fallacy is false.

Hubbard and Evans concluded that the combination of ease of use, with difficulty in tracking results, have contributed to the proliferation of scoring methods (Hubbard & Evans, 2010).

2.2 DoD Guidelines for Risk Analysis

The DoD RIO guidebook seeks to establish the quality thinking and judgment applicable for all acquisition programs throughout their lifecycle (DASDSE, 2017). For a typical acquisition program, risk reduction activities are planned at the earliest and continued throughout the program following the DoD Acquisition Lifecycle process shown in Figure 4.

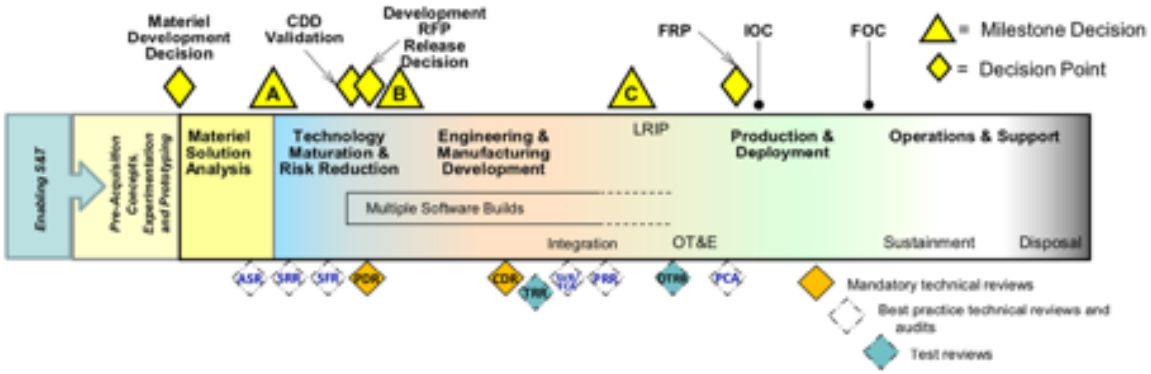


Figure 4 - DoD Acquisition Life Cycle (DASDSE, 2017)

The guide is designed to produce risk mitigation plans, which provide the substantive steps a program will take to mitigate its actual risks (DASDSE, 2017). Figure 5 illustrates the risk analysis phase, which is part of the DoD Risk Management Five-Step Process. In this phase, practitioners are asked to do three main tasks:

- Evaluate the probability for the risk event,
- Evaluate the consequence of the risk event, and
- Determine the RE and mitigation strategy.

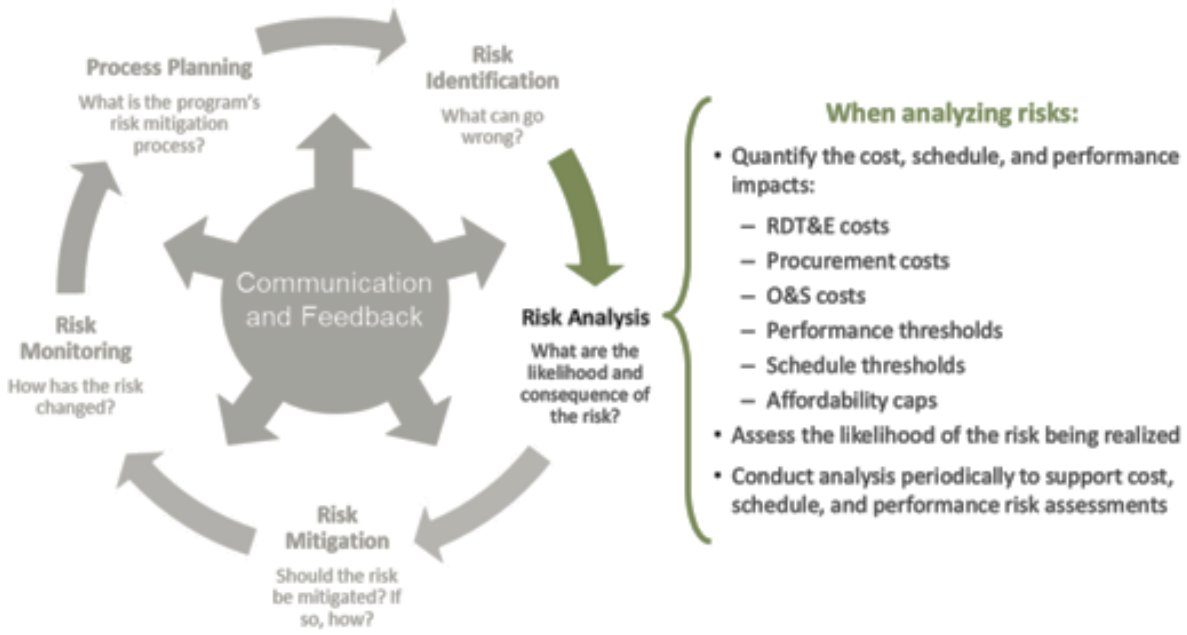


Figure 5 - DoD Risk Analysis Phase (DASDSE, 2017)

This overview focuses only on the risk analysis section 3.3 of the guide and will consider that it is not the only guidance in the department for risk management. However, many of the other available guides, like the internal instructions, recommend a similar path for scoring risks. When estimating consequence, the initial guidance is to tailor the criteria based on program-specific circumstances and use Table 6 to determine the consequence based on a given set of assumptions. This can be particularly challenging for projects that are of the research and development (R&D) nature in which earlier historical information such as cost and technical performance is close to nil.

Table 6 - Typical Consequence Table from DoD RIO guidebook (DASDSE, 2017)

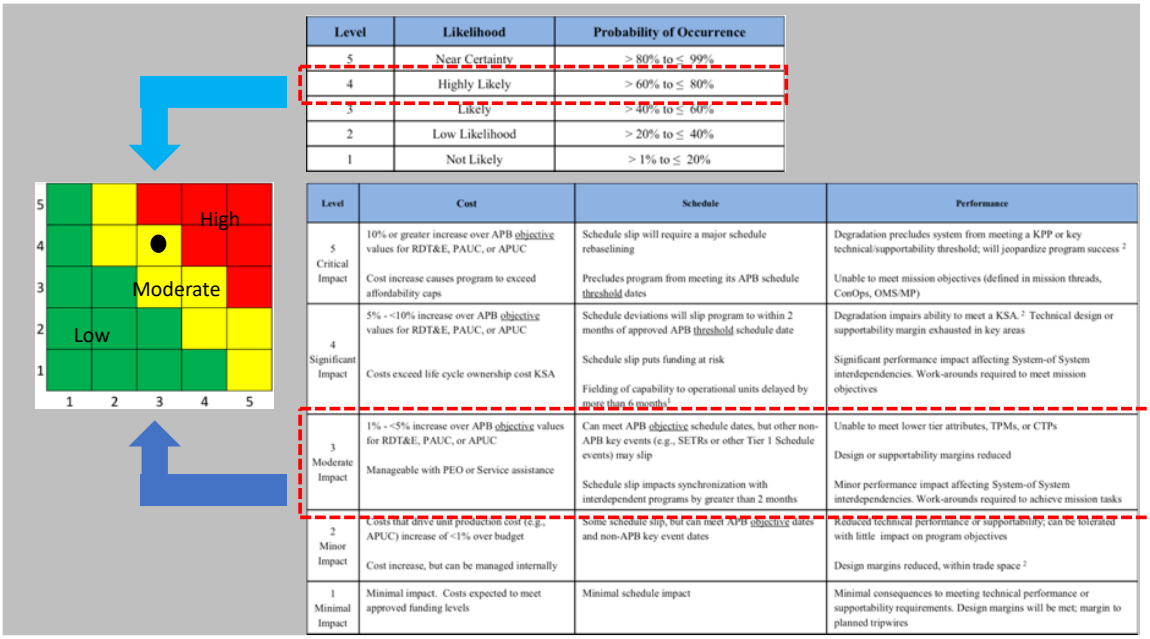
LEVEL	COST	SCHEDULE	PERFORMANCE
5 Critical Impact	10% or greater increase over APB objective values for RDT&E, PAUC, or APUC. Cost increase causes program to exceed affordability caps.	Schedule slip will require a major schedule re-baselining. Precludes program from meeting its APB schedule threshold dates.	Degradation precludes system from meeting a KPP or key technical/supportability threshold; will jeopardize program success. Unable to meet mission objectives (defined in mission threads, ConOps, OMS/MP).
4 Significant Impact	5% - <10% increase over APB objective values for RDT&E, PAUC, or APUC. Costs exceed life cycle ownership cost KSA.	Schedule deviations will slip program to within 2 months of approved APB threshold schedule date Schedule slip puts funding at risk. Fielding of capability to operational units delayed by more than 6 months.	Degradation impairs ability to meet a KSA. Technical design or supportability margin exhausted in key areas. Significant performance impact affecting System-of System interdependencies. Work-arounds required to meet mission objectives.
3 Moderate Impact	1% - <5% increase over APB objective values for RDT&E, PAUC, or APUC. Manageable with PEO or Service assistance.	Can meet APB objective schedule dates, but other non- APB key events (e.g., SETRs or other Tier 1 Schedule events) may slip. Schedule slip impacts synchronization with interdependent programs by greater than 2 months.	Unable to meet lower tier attributes, TPMs, or CTPs. Design or supportability margins reduced. Minor performance impact affecting System-of System interdependencies. Work-arounds required to achieve mission tasks.
2 Minor Impact	Costs that drive unit production cost (e.g., APUC) increase of <1% over budget. Cost increase, but can be managed internally.	Some schedule slip, but can meet APB objective dates and non-APB key event dates.	Reduced technical performance or supportability; can be tolerated with little impact on program objectives. Design margins reduced, within trade space.
1 Minimal Impact	Minimal impact. Costs expected to meet approved funding levels.	Minimal schedule impact.	Minimal consequences of meeting technical performance or supportability requirements. Design margins will be met; margin to planned tripwires.

When estimating likelihood, the interesting part is that it asks the practitioner to provide a well-defined risk event or condition to assess against Table 7, with almost all arbitrary defined ranges, values, and definitions.

Table 7 - Typical Likelihood Table from DoD RIO guidebook (DASDSE, 2017)

Level	Likelihood	Probability of Occurrence
5	Near Certain	>80% to ≤ 99%
4	Highly Likely	>60% to ≤ 80%
3	Likely	>40% to ≤ 60%
2	Low Likelihood	>20% to ≤ 40%
1	Not Likely	>1% to ≤ 20%

However, depending on the program resources, time spent, or costs, the practitioner may choose to use Monte Carlo methods that are rarely considered because of the perception that this approach is not cost-effective (Pate-Cornell, 2002). Once the practitioner has established the likelihood and consequence, then the DoD RIO guidebook describes the reporting process, which, in this case, is the risk matrix. The matrix is subdivided into twenty-five different sections and colored from green, yellow and red. The color scale, Low (green), Moderate (yellow) and High (red), prioritize each assigned risk. Using the 5x5 matrix, if the likelihood was scored at three and a consequence was scored at three, then the RE would fall to a Moderate scale as Figure 6 illustrates.



Level	Likelihood	Probability of Occurrence
5	Near Certainty	> 80% to ≤ 99%
4	Highly Likely	> 60% to ≤ 80%
3	Likely	> 40% to ≤ 60%
2	Low Likelihood	> 20% to ≤ 40%
1	Not Likely	> 1% to ≤ 20%

Level	Cost	Schedule	Performance
5 Critical Impact	10% or greater increase over APB objective values for RDT&E, PAUC, or APUC Cost increase causes program to exceed affordability caps	Schedule slip will require a major schedule rebaselining Precludes program from meeting its APB schedule threshold dates	Degradation precludes system from meeting a KPP or key technical/supportability threshold; will jeopardize program success [2] Unable to meet mission objectives (defined in mission threads, ConOps, OMS/MP)
4 Significant Impact	5% - <10% increase over APB objective values for RDT&E, PAUC, or APUC Costs exceed life cycle ownership cost KSA	Schedule deviations will slip program to within 2 months of approved APB threshold schedule date Schedule slip puts funding at risk Fielding of capability to operational units delayed by more than 6 months[1]	Degradation impairs ability to meet a KSA. [2] Technical design or supportability margin exhausted in key areas Significant performance impact affecting System-of System interdependencies. Work-arounds required to meet mission objectives
3 Moderate Impact	1% - <5% increase over APB objective values for RDT&E, PAUC, or APUC Manageable with PEO or Service assistance	Can meet APB objective schedule dates, but other non-APB key events (e.g., SETRs or other Tier 1 Schedule events) may slip Schedule slip impacts synchronization with interdependent programs by greater than 2 months	Unable to meet lower tier attributes, TPMs, or CTPs Design or supportability margins reduced Minor performance impact affecting System-of System interdependencies. Work-arounds required to achieve mission tasks
2 Minor Impact	Cross-link drive unit production cost (e.g., APUC) increase of <1% over budget Cost increase, but can be managed internally	Some schedule slip, but can meet APB objective dates and non-APB key event dates	Reduced technical performance or supportability, can be tolerated with little impact on program objectives Design margins reduced, within trade space [2]
1 Minimal Impact	Minimal impact. Costs expected to meet approved funding levels	Minimal schedule impact	Minimal consequences to meeting technical performance or supportability requirements. Design margins will be met; margin to planned tripwires

Figure 6 - Typical Risk Reporting from DoD RIO guidebook (DASDSE, 2017)

The guide does not provide the practitioner with much else in the form of methodological guidance, such as what tools to use, or what practical method to use when no data is available or how to analyze data sets. Most practitioners are left to assume that this is the only risk analysis process. The DoD RIO guidebook does provide an example of how to prioritize high-level risks based on the cost impact, but it advises that technical performance impact should be prioritized above lower priorities. Table 8 is a risk prioritization example from the DoD RIO guidebook.

Table 8 - Risk Prioritization Example

Risk	Likelihood	Consequence Cost	Risk-Weighted Consequence	Cost to Mitigate	Expected Return on Investment
Risk 1	20%	$10M	$2M	$1M	$1M (1:1)

Risk 2	70%	$10M	$7M	$1M	$6M (6:1)
Risk 3	40%	$36M	$14.4M	$2M	$12.4M (6:1)
Risk 4	60%	$5M	$3M	$.5M	$2.5M (5:1)
Total		**$61M**	**$21M**	**$4.5M**	

The source claims that, by using scoring methods, practitioners are following best practices; however, there is a significant shortcoming that relates to the scope and method given to these practitioners. Additional methods could supplement and enhance the scope to increase the technical rigor, and value to DoD programs. Furthermore, the scope is tailored primarily to a program of record (PoR) project. PoR projects are usually synonym with a project that has been officially approved as a development program in the government and carry approved future year funding. The DoD RIO guidebook scope makes no adjustments nor provisions for R&D and rapid prototype projects for which the risk management process is less forgiving due to the projects' shorter timeline.

2.3 Engineering Management Body of Knowledge

Domain five of the Engineering Management Body of Knowledge (EMBOK) relates to project management. As defined by the PMBOK a project is a temporary endeavor to create a product or service eventually has to address risks (Project Management Institute, Inc., 2017). The EMBOK follows much of the Project Management Book of Knowledge (PMBOK) guidance as in both the EMBOK and PMBOK; risk management planning starts early in the project's lifecycle. Similar to the DoD RIO guidebook, engineering management defines a four-step process to address an uncertain future event or condition (Shah & Nowocin, 2015). In the EMBOK, risks are defined as future events or conditions that, if they occur, have a positive or negative effect on the project (Shah & Nowocin,

2015). In the EMBOK, just like the DoD RIO guidebook, the process assigns three main tasks:

- Evaluate the probability for the risk event,
- Evaluate the consequence of the risk event, and
- Determine the RE and mitigation strategy.

Once a risk is identified, the practitioner assesses the RE in terms of threats or opportunities based on the Probability and Impact Matrix depicted in Figure 7.

Probability and Impact Matrix

Probability	Threats					Opportunities				
0.90	0.05	0.09	0.18	0.36	0.72	0.72	0.36	0.18	0.09	0.05
0.70	0.04	0.07	0.14	0.28	0.56	0.56	0.28	0.14	0.07	0.04
0.50	0.03	0.05	0.10	0.20	0.40	0.40	0.20	0.10	0.05	0.03
0.30	0.02	0.03	0.06	0.12	0.24	0.24	0.12	0.06	0.03	0.02
0.10	0.01	0.01	0.02	0.04	0.08	0.08	0.04	0.02	0.01	0.01
	0.05	0.10	0.20	0.40	0.80	0.80	0.40	0.20	0.10	0.05

Impact (numerical scale) on an objective (e.g., cost, time, scope or quality)

Each risk is rated on its probability of occurring and impact on an objective if it does occur. The organization's thresholds for low, moderate or high risks are shown in the matrix and determine whether the risk is scored as high, moderate or low for that objective.

Figure 7 - EMBOK Probability/Impact Matrix

Much like the DoD guidance, the EMBOK risk assessment method is an ordinal scale scoring method; therefore, as other researchers have suggested it can suffer from the same subjective inputs, biases, and flaws.

2.4 Overview of Cooke's Classical Method

Probabilistic thinking is the cornerstone of any decision-making method. Cooke elaborates on the subject by describing the importance of logical thinking when assessing uncertainty in probabilities and the challenges most people experience (Cooke R. M., 1991). Fundamental mistakes are often among the simplest when evaluating logical statements. With the advancement of technology, computer-aided decision-making is transitioning and improving.

Due to technology maturation, one problem that computer-aided decision-making has faced in the past is the determining conditional probabilities certainty factors, as these do not meet some of the probabilities axioms they should (Cooke R. M., 1991). The SEJ method is invoked when there is a high level of uncertainty in the decision-making process, a lack of empirical data, or absence of a simple consensus among the experts (Cooke & Goossens, 2008). After all, the goal of SEJ is to reach some rational consensus among the experts. This rational consensus refers to a decision-making process on how the uncertainty will be characterized by the experts. Alternatively, the article, "Quantifying Scientific Uncertainty from Expert Judgement Elicitation" written by Cooke and Aspinall describes the goal of the SEJ method to be to quantify uncertainty, rather than to remove it, from a decision process (Cooke & Aspinall, 2011).

When estimating uncertainty in probabilities using a decision-making process such as SEJ, the practitioner must account for the degree of belief to the extent the expert gives the approximation. These heuristics can often lead to biases and errors in the subjective assessment. Cooke presents four types of heuristics affecting an SEJ: availability, anchoring, representativeness, and control (Cooke R. M., 1991). The availability of

heuristics can be best described as a mental shortcut. The information is recalled from recent events and given priority to any critical judgment. When describing the anchoring heuristics, it is often the case that practitioners are estimating probabilities and fixating on these values and adjusting or correcting afterward. In instances when practitioners are evaluating the probabilities for two events, they often assess the resulting probabilities with a degree of similarity between events. It is this comparison that leads to the representativeness heuristics. Lastly, control heuristics are when practitioners think they can influence an event outcome, but they cannot. Also, when assessing these probabilities, it is essential to consider if the expert is biased for overconfidence or if overconfidence is expressed as poor calibration (Cooke R. M., 1991). Perhaps common among biases are the conjunction and base rate fallacies. The conjunction fallacy can be best described by what researchers call the "Linda Problem," experiment.

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and participated in anti-nuclear demonstrations* (Kahneman & Tversky, 1983).

- Which is more probable?
  - Linda is a bank teller.
  - Linda is a bank teller and is active in the feminist movement.

According to the research, in most cases, the second option is rated as the most likely answer, but that is incorrect. It cannot be more likely that Linda is a bank teller and a feminist than she is a bank teller (Kahneman & Tversky, 1983). This being the conjunction fallacy. The base rate fallacy applies to the assessment of probabilities conditional upon

the prior information. If practitioners were to change the original assessment and ignore new information, then the resulting likelihood is inappropriate and treated as base rate neglect due to the information omitted (O'Hagan & et.al., 2006). Also, as Cooke explains, this bias is generally present when experts tend to ignore prior information or available data for a specific case or scenario (Cooke R. M., 1991). Cooke explains that the base rate fallacy often leads to overconfidence and poor calibration (Cooke R. M., 1991). The SEJ treats the information elicited from the experts as empirical data and uses the scientific method and the following fundamental principles to enable rational consensus:

- Accountability – If needed, all data, evaluations, and responses from the experts must be peer-reviewed and reproducible with the goal of getting the same outcome.
- Empirical Control – Expert assessments are subject to empirical quality controls, same as any other quantitative research or data analysis facts gathering.
- Neutrality – The method for evaluating experts should encourage their actual opinions and must not bias the results.
- Fairness – Experts are not pre-judged by the analysis nor is the process in which the experts provide the elicited data to be pre-judged.

Early structured elicitation methods such as the Delphi method, suggests that any consensus achieved is due mainly to group conformity pressure or mediated by the most influential group participant (Cooke & Aspinall, 2011). Alternatively, the *equal weighting method* is considered one of the simplest aggregations for expert elicitations. Equal weight combinations are usually statistically acceptable but not always because there has been statistical finding in which it can be rejected. According to research, in approximately 15%

of the studies examined, the hypothesis that the equal weight combination was statistically accurate would be rejected at the 5% level (Cooke & Aspinall, 2011).

The *performance weighting method* which is usually characterized as a linear pooling in which weights are assigned to experts based on their individual assessment performance. The *classical model* employs this performance weight method and the fundamental principles to obtain an uncertain judgment from the experts, which relies on proper scoring when combining the performance weights of the combined individual judgments into a rational consensus that can ultimately inform the decision-maker (Cooke & Goossens, 1999). These scores allow the analyst to 1) quantify an expert's performance based on the subjective uncertain judgment and compare to the other experts, 2) create performance distributions for all the experts and 3) evaluate the combined expert performance in achieving a rational consensus.

According to Clemen, Cooke's classical method is an example of a ''linear opinion pool,'' a weighted arithmetic average of the experts' probability distributions (Clemen, 2008). It is characteristic of weighting the expert's relative performance using scoring rules. Cooke explains that the classic method's scoring rule is for the expert's average probabilities (Cooke R. M., 1991). Both Aspinall and Clemen documented their concerns regarding the possibility of experts trying to game the elicitation phase by providing less-than-honest answers (Aspinall W. P., 2006) (Clemen, 2008). In this instance, an expert could get better than average calibration scores, but the information scores will be lower, given the less-than-honest answers. Ultimately, the overall score (weight) for this expert would be reduced.

While practitioners of the classical method have the option to perform performance weight (PW) or equal weight (EW) combination on calibration variables, these can be tested using in-sample and out-of-sample and produce a resultant calibration and information scores. According to Cooke, PW is superior when doing out-sample testing because it delivers statistically acceptable calibration and information scoring for the expert's results (Cooke R. M., 2015). When used on calibration variables, EW combination provides good calibration scoring at the expense of low information scoring. Research completed by Eggstaff and et al. confirms that cross-validation of different methods has shown that out-of-sample testing also prefers PW (Eggstaff, Mazzuchi, & Sarkani, 2014).

2.4.1 Assumptions

In the case of this research study, to be able to accept the empirical contributions from an expert, there must be two qualifiers met, or, in this case, two assumptions:

- The volunteer providing the subjective assessment is proficient in a relevant subject(s).
- The volunteer providing the assessment understands the basic principles of probability and uncertainty.

These qualifiers will afford the SEJ process and the experts the empirical control and contributions needed during the elicitation process.

2.4.2 Limitations

According to Cooke and Aspinall, large-scale SEJ can become costly due to the required number of experts and personnel to develop the seed variables (Cooke & Aspinall, 2011). Also, the performance and equal weight criteria could be improved in this situation, due to being resource intensive; however, Cooke and Goossens observe that the most

significant limitation regardless SEJ type is the assumption of independence in uncertainty analysis, which is not unique to this method (Cooke & Goossens, 2008). It is also necessary to recognize each expert's limitations, expertise, and proficiencies when judging probability and uncertainty in a decision-making environment (Cooke & Aspinall, 2011).

2.5 An Overview of the Bayesian Inference Method

BIs require that the practitioner provide the prior probability and process expert assessments by evaluating the Bayes' Theorem (Cooke R. M., 1991). BI methods have been around since the 1700s; Pierre-Simon Laplace, one of the first champions of BI methods, performed the quantitative inference calculation using Bayes' methods. Currently available risk analysis approaches can be categorized as either being frequency-based or Bayesian-based approach (Aven & Heide, 2009). In a frequency-based approach, probabilities are estimated based on the relative occurrence of the events it portrays. This research study does not focus on the frequency-based approach; instead, it explores how the Bayesian approach could improve probability assessments. The former is defined by how current and future events and outcomes provide a measure of uncertainty in a probability. NASA currently uses a Bayesian-based approach, known as Bayesian inference for all the space vehicle safety checks and launch capabilities to determine their risk profile (Stamatelatos & Dezfuli, 2011). NASA has achieved essential milestones in probabilistic risk assessment by enhancing the decision-making process and incorporating this tool into their daily mission and safety goals.

The inference process used starts with an understanding of three key terms: data, information, and knowledge (Kelly & Smith, 2011). Data is an observed value of a physical process, as information, it is the result of processing and collecting data. Thus,

knowledge is the collection of information or a result of processing data. When evaluating an inference, there are two fundamental types of models aleatory and deterministic. Aleatory models primarily deal with probabilities, and deterministic models mainly deal with relations, such as equations. Given these parameters, one must wonder why Probabilistic Risk Assessments (PRA) are not employed more often in risk analysis. As NASA has shown, PRA is an exceedingly reliable tool.

Aven and Heide investigate the reliability and validity of two different types of Bayesian approaches, among other methods to determine if risk analysis met the scientific requirements established below (Aven & Heide, 2009):

- The scientific work shall comply with known methods and processes to ensure it can be reproduced.
- The analysis is relevant and useful in ensuring it contributes to the field by solving a problem.
- The analysis results are reliable and valid.

Aven and Heide conclude that Bayesian approaches that predicted observable parameters, similar to BI, do not meet reliability criteria established in their paper unless the data and methods are fixed (Aven & Heide, 2009).

Other studies explore the practicality of probabilistic risk assessments from the cost perspective and discuss several interesting points starting with familiarity and education. According to one of these studies, there is a challenge when it comes to finding an adequate population of practitioners and decision-makers that are both experienced and proficient in PRA tools (Lund, 2008). The implementation of any good model and solution is critical since it hinges on both the practitioner's and the decision-maker's experience and

proficiency to succeed. The other essential element eluded to is the cost (Lund, 2008). In essence, practitioners need to determine the risk impact's cost before engaging a full-blown model. The study recommends a decision matrix (refer to Figure 8) when the PRA model becomes cost prohibitive (Lund, 2008).

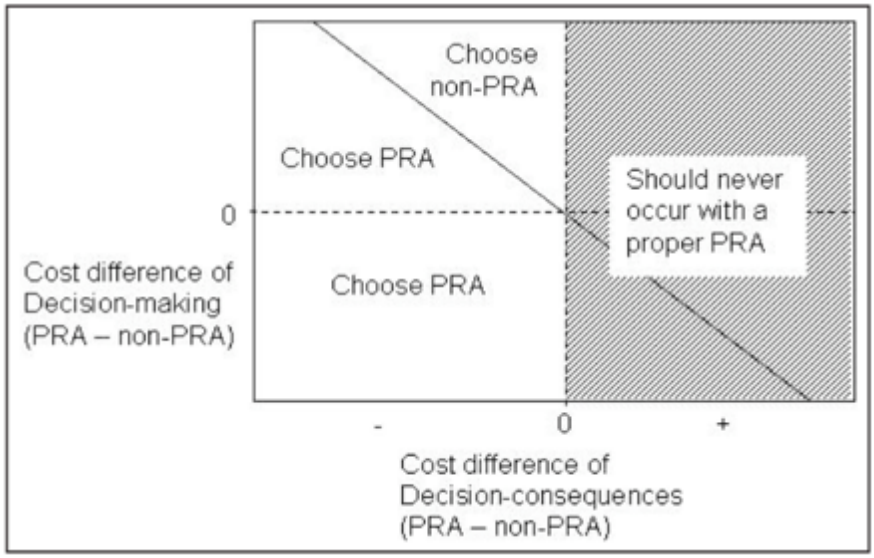


Figure 8 - PRA Cost Decision Matrix (Lund, 2008)

In his concluding statements, Lund mentions that any PRA model, even if it is somewhat data starved, will yield better results for the risk analysis. Cooke concludes that, as with the classical model, successful implementation of a Bayesian inference depends on how well the relevant variables are defined by the analyst (Cooke R. M., 1991).

2.5.1 Assumptions

One of the first assumptions the BI makes is that there is data available for the modeling and the second assumption that the data available characterizes the problem. If the data was not mined from historical archives or collected from test trials or determined from a root cause analysis, then the model does not and cannot determine the difference of good quality data. Data gaps and uncertainty can add to the model, but the quality of the data

and its representation to the problem is a critical assumption of the BI. If the practitioner fails to understand this assumption, the BI will not yield the expected results.

2.5.2 Limitations

Perhaps ten or fifteen years ago, the computational demands of performing a Markov Chain Monte Carlo (MCMC) algorithm to determine the BI would have been significant. Fortunately, this limitation has been overcome with the aid of software tools like WinBUGS, OpenBUGS, and many others, in addition to the advanced in computational performance of today's CPUs and GPUs. While developing the model, one can determine to use a conjugate prior based on the type of data and problem being solved or use a Jeffrey's prior if there is a shortage of data (or significant uncertainty). The limitations of using Jeffrey's prior in some cases is that the model can become too conservative and depends on the model used it could cause the posterior distribution to become improper. Moreover, the implementation of any good model and solution hinges on both the practitioner's and decision-maker's experience and proficiency to succeed.

Chapter 3 Research Methodology

This chapter provides an overview of the decision-making process used when establishing a tailored SEJ and BI methodology within the basic risk model in the context of the DoD risk analysis phase. It elaborates on the classical method steps as part of the expert elicitation and expands on the BI data analysis efforts to improve probability estimates.

3.1 Decision-Making Process Overview

In their paper, "On the Quantitative Definition of Risk" Kaplan and Garrick defined risk analysis as a series of three questions: What can happen? How likely is to happen? So, if it does happen, what are the consequences? (Kaplan & Garrick, 1991). It is this question, how likely is the risk to occur that this research study will focus. As described in the introduction the DoD risk management process directs the user towards a series of steps that ultimately ends in the use of a scoring system with little to no guidance in the use of elicitation or quantitative methods *established on quantitative programmatic and engineering analyses to the extent practicable* (DASDSE, 2017). Given this process improvement opportunity the selected methodology, Structured Expert Judgment (SEJ) is the method used when subjective probabilities are employed in the basic risk model. The Bayesian inference (BI) method analysis efforts to improve probability estimates will be explored in the expanded basic risk model which is depicted in Figure 9.

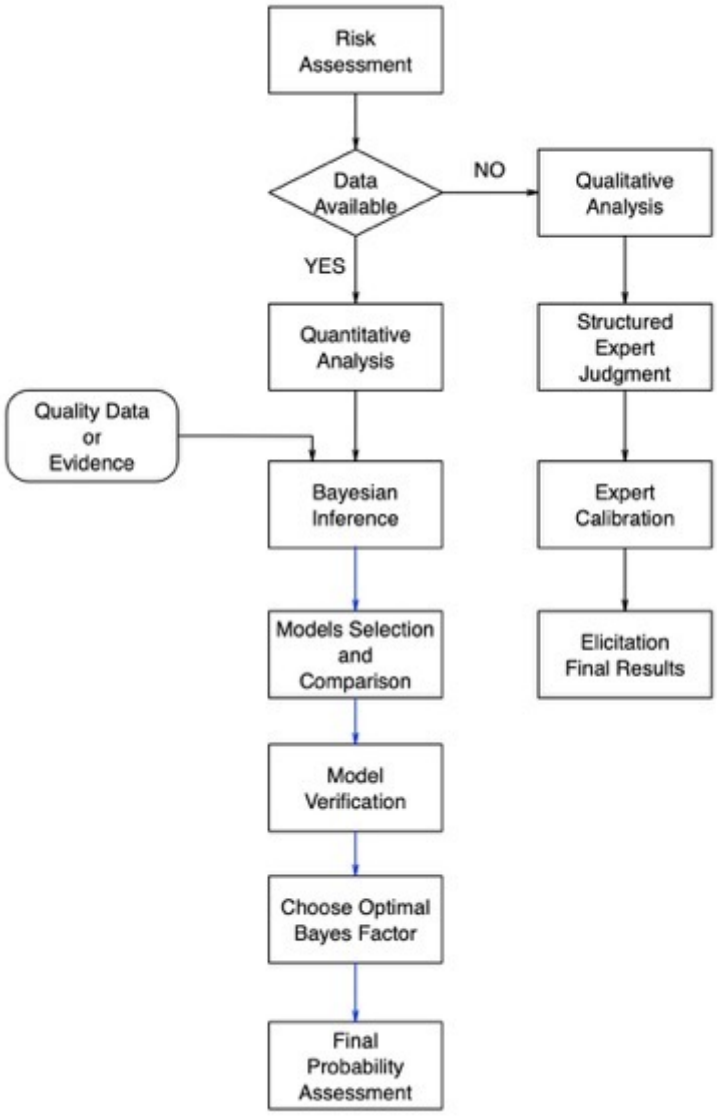


Figure 9 - Basic Risk Model (Expanded)

In his 1997 paper, "The Words of Risk Analysis" Kaplan describes the evidence-based approach a subset of the body of evidence, as depicted in Figure 10, as the means to use probability theory (Kaplan, 1997). The suggested BI methodology can be used in the objective portion of the risk model which will be explained in more details in section 3.3.

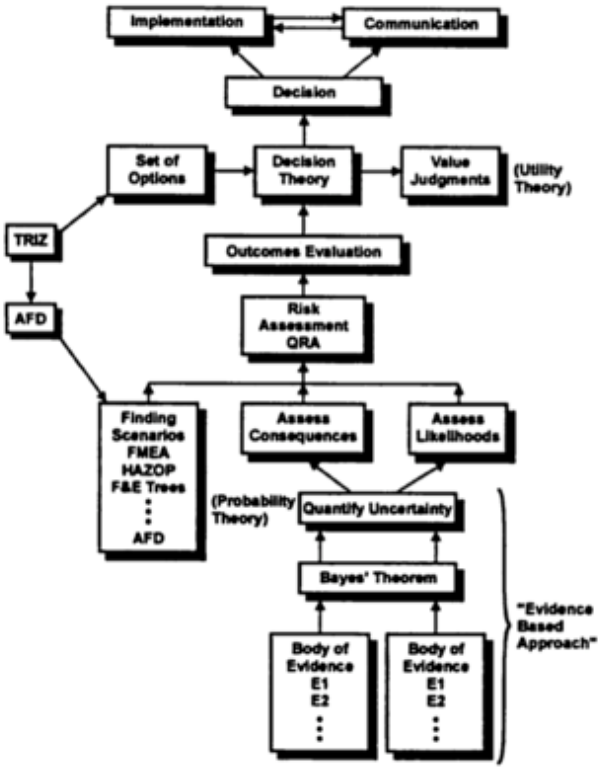


Figure 10 – Body of Evidence Approach (Kaplan, 1997)

3.2 Tailored Elicitation (Based on Classical Method)

This section will describe the steps leading to structured elicitation method described in the Cooke's classical method, but by adjusting the scope, expert selection process and harmonizing the process to the user, the resulting method was tailored from the original classical approach. The following sections describe how the tailored methodology was implemented.

3.2.1 Pre-Calibration

3.2.1.1 Obtaining Project Data

It was agreed in advanced that any data donated by active projects to this research study would have the project details anonymized. Once the data was secured, the process of

reviewing it and categorization began. Data classification is explained in the following section. Three projects donated the raw data for the SEJ method. Table 9 below summarizes the data sources.

Table 9 - SEJ Raw Data Sources (Summary)

Project	Data Type	Risks Identified
P1	Subjective	158
P2	Subjective	36
P3	Subjective	25

3.2.1.2 Data Review and Categorization

Data review and classification was an essential step in the SEJ method since it allowed the research study to determine which risks were used for the survey questions. The selection process considered the following attributes from the current project databases:

- SEJ principle for expert's knowledge pool.
- Ease to adapt risk item to sponsor's sharing terms of anonymity.

Based on the above categorization and using data from the donating projects as shown in Table 9 as the data source, the following seed variables within the project risks were chosen with their respective realization values as depicted in Table 10.

Table 10 - Calibration Phase Seed Variables with Realizations

No.	Seed Variables	**Realization** (In terms of Risk Exposure)
1	**Finish on Time** - Uncertainty to complete specific deliverable in each timetable	13
2	**Adapter Quality** - Uncertainty of adapter quality after completing specific tests	23

3	**Funding Award** - Task funding uncertainty given funding oversight authority	14
4	**Capability Implementation** - Hardware capability uncertainty based on scope changes	12
5	**Draft RMF Requirement** - Timetable uncertainty to implement process requirements	17
6	**On-Time Delivery of Component to Sponsor** - System integration and deliverable uncertainty based on requirements and timetable	17
7	**Prevent Funding Shortage** - Funding uncertainty in a configuration control and requirement establishment	22
8	**Placing Raw Materials Orders** - Funding uncertainty for procurement of resources	19
9	**Program Requirement** - Documentation and requirements traceability given uncertain resource allocations	17
10	**Model Calibration** - Uncertainty in a model calibration given poor historical data	13

A similar process was taken when reviewing and categorizing the project data for the elicitation phase of the study. The categorization and data from the donating projects shown in Table 9 are used as data sources to determine the *variables of interest* for this phase. The following target variables were identified, as depicted in Table 11. The term target variable and variable of interest is used interchangeably in this praxis.

Table 11 - Elicitation Phase Target Variables (*Variables of Interest*)

No.	Elicitation (Variables of Interest) – With No Realization
1	**Widget Malfunction** - Typically, the widget functions its best mechanical functional design parameters by moving from an open and closed position. Assuming the widget material is aluminum. Its normal thermal loading capacity is 1000 cycles at room temperature.
2	**Reliability** - Theoretically defined as 1 - "probability of failure;" or, in terms of availability, {uptime/ (uptime + downtime)}
3	**Communication** - Leadership misinformation may affect communication and increase the level of uncertainty.
4	**Technical Performance** - A prediction and demonstration of the degree of an expected or actual achievement of selected technical goals.

3.2.1.3 Experts Selection

Requests for volunteers for the research study was elicited via email. A network of highly qualified engineers and scientists already exposed to Projects P1, P2, P3, and P4 were pooled. The resultant participant's list is coded for IRB purposes and to protect the neutrality and fairness principles of the SEJ method. Table 12 depicts the final tally of volunteers.

Table 12 - Research Study Experts (Spartan Team)

N	Experts
1	S117
2	S118
3	S119
4	S120
5	S121
6	S122
7	S123
8	S124
9	S125
10	S126
11	S127
12	S128
13	S129
14	S130
15	S131
16	S132

3.2.1.4 Initial Survey and Consent

During the first session, participants were presented with a consent form and an initial survey as depicted in Appendix C. The consent form explained the necessary qualifications, the goals of the research, the process for which the SEJ will be conducted and finally the risks the participants might be exposed. The acceptance process was simple. If the participants agreed to all the terms in the consent form, then they were allowed to participate in the study. The initial survey primary intent was to get a general sense of all

participants' field of expertise as well as their pay grade code. The initial survey information will later inform the calibration phase and aid to prove one of the hypotheses. Once the participants completed the initial survey, the student investigator was able to classify them as depicted in Table 13. Each grade code AZ, BZ, and NX represent a different engineering overhead expenditure per hour (EO/h).

Table 13 - Spartan Team w/ Pay Grade Classification

N	Experts	Grade
1	S117	NX
2	S118	BZ
3	S119	NX
4	S120	BZ
5	S121	AZ
6	S122	BZ
7	S123	BZ
8	S124	BZ
9	S125	BZ
10	S126	NX
11	S127	BZ
12	S128	BZ
13	S129	BZ
14	S130	NX
15	S131	BZ
16	S132	BZ

Once consent was received by all participants, and the initial survey was completed then the preparation for the calibration phase started.

3.2.2 Calibration Phase

3.2.2.1 Expert's Training

During the expert's training phase, volunteers were trained in the concepts of the Cooke's Classical Method principles and given a basic understanding of the probabilistic and statistical foundations to understand the method. The training initially consisted of a

nine (9) weeks curriculum, but it expanded to twelve (12) due to schedule and other conflicts. Among the topics discussed during the training sessions were:

- A basic definition of a Structured Elicitation
- A basic definition of a Subjective Probability
- Principles of a Structured Elicitation (Accountability, Empirical Control, Neutrality, and Fairness)
- The basis of a case structure, target, and query variables
- Calibration and Information Scores

3.2.2.1.1 Defining Structured Elicitation

One of the primary goals of a formal elicitation of expert judgment is to remove as much subjectivity from decision-making as possible. SEJ is characterized when the information is elicited, codified and combined with the knowledge of people. It is appropriate and justified to use SEJ when quantitative data is missing, and the quality of the data is dubious. As well as if obtaining the quantitative data by statistical means is not possible (Cooke & Aspinall, 2011). It is important to note that in a decision-making process the uncertainty in the problem can be addressed by the use of models and predictive tools (Cooke & Goossens, 1999).

3.2.2.1.2 Defining Subjective Probability

Someone's degree of belief based on an uncertainty value determines a subjective probability (O'Hagan & et.al., 2006). These depend on the knowledge the person has and are often bias towards their experiences, emotions and even their memories when evaluating a subjective probability.

3.2.2.1.3 Principles of a Structured Elicitation

The following fundamental principles or conditions are needed for every SEJ to be conducted adequately (Cooke & Aspinall, 2011). These are:

- Accountability: All data, assessments, and processing tools are open to peer review and results must be reproducible by competent reviewers.
- Empirical control: Quantitative expert assessments are subjected to empirical quality controls.
- Neutrality: The method for combining/evaluating expert opinion should encourage experts to state their actual opinions and must not bias results.
- Fairness: Experts are not pre-judged, before processing the results of their assessments.

3.2.2.1.4 Tailoring of Case Structure

The process of tailoring the SEJ method consisted primarily in presenting baseline problems representatives from Projects P1, P2, and P3. To that end, the case structure for each is design to address the problem statement of the projects but at the same time masks the sensitive aspects of the projects that did not pertain to this research study. The case structure was tailored as follows:

- The case structure – defines the problem statement from data sources drawn from the donating projects as depicted in Table 9. For the calibration phase, the case requests from the experts to evaluate the uncertainty of a seed variable against its realization value, see Table 10 for reference. For the elicitation phase, the case asks from the experts to assess uncertainty based on the variables of interests, see Table 11 for reference.

3.2.2.1.5 Classical Model Performance Measures

Using the guidance from Cooke, Colson and Goossens and some of the same methodology equations from Eggstaff and et al. these were the steps followed to establish the necessary calibration and informational foundation on the SEJ method (Cooke & Goossens, 1999) (Eggstaff, Mazzuchi, & Sarkani, 2014). Some of the basic terminology for the equations to follow and primary variables used for the methodology are defined below:

- Total number of experts in the research study is defined by variable, *E*.
- A singular expert is defined by variable *e*.
- The set of calibration variables to be assessed is denoted by, $\{X_1, \dots, X_n\}$.
- The number of seed variables being evaluated in a particular probability bin is denoted by *n*.
- A singular calibration question is defined by the variable *i*.
- The set of true realization values in a calibration variable is denoted by, $\{x_1, \dots, x_n\}$.
- The quartile or percentile bins are given in 5%, 50%, and 95% buckets or as denoted by $\left(x_{5,i,e}, x_{50,i,e}, x_{95,i,e}\right)$.

Where the 5, 50, and 95 are the selected quartile percentile split used in the research study, the sample distributions of the expert's inter-quartile intervals as depicted in Equation 1, can be defined as:

Equation 1 - Expert's Inter-Quartile Intervals

$$I_{1,i,e} \stackrel{\text{def}}{=} \left[a_i, x_{5,i,e}\right]$$

$$I_{2,i,e} \stackrel{\text{def}}{=} \left[x_{5,i,e}, x_{50,i,e}\right]$$

$$I_{3,i,e} \stackrel{\text{def}}{=} \left[x_{50,i,e}, x_{95,i,e}\right]$$

$$I_{4,i,e} \stackrel{\text{def}}{=} \left[x_{95,i,e}, b_i\right]$$

Where, [a, b] is the intrinsic range defined given the 5, 50, and 95 quartile percentiles split used. In this case intrinsic range is between 1 and 25. The next step is to define the observed bin variables based on the seed variables and responses from the experts. This variable as depicted in Equation 2, is described as:

Equation 2 - Seed variables frequency

$$s_i = \frac{Observed\ seed\ value\ in\ the\ quantile\ bin}{n}$$

With the above ground work explained, the foundational hypothesis for a well-calibrated expert-hypothesis can be defined. The expert-hypothesis will be instrumental in determining which experts are calibrated based on the *p-value* using empirical control and empirical data through the SEJ process. As independence is not an assumption about expert's distribution during the elicitation, but a need to solve the problem (Colson & Cooke, 2017). The decision-maker (DM) scores the experts as the statistical likelihood of the following expert-hypothesis (H_e) (Colson & Cooke, 2017):

H_e: *"the inter-quantile interval containing the true value for each variable is drawn independently."*

It is important to emphasize that the SEJ method is not rejecting H_e; rather it measures the degree by which the data support H_e that the expert assessments are correct (Colson &

Cooke, 2017). Each of the sample distributions there is an independent *p* distribution as depicted in Equation 3 where:

Equation 3 - Probability distribution

$$p = (p_1, p_2, p_3, p_4) = (0.05, 0.45, 0.45, 0.05)$$

In summary, if the values given by each expert *e* are independent of the distributions *I* then the sample distribution as depicted in Equation 4 can be written in the following form:

Equation 4 - Sample Distribution Equation

$$2NI(s,p) = 2N \sum_{i=1,\ldots,4} s_i \ln\left(\frac{s_i}{p_i}\right)$$

(Eggstaff, Mazzuchi, & Sarkani, 2014)

Where *I(s,p)* is the function of measure of relative information. If there is a sufficiently large number of calibration variables *N*, then Equation 4 above becomes the calibration score as depicted in Equation 5. The calibration score resembles a chi-square distribution with three degrees of freedom (one for each quantile).

Equation 5 - Calibration Score Equation

$$cal(e) = 1 - \chi^2_{(3)}\big(2NI(s,p)\big)$$

Where: $\chi^2_{(3)}\big(2NI(s,p)\big)$ is the $\Pr\{\chi^2_{(3)} < 2NI(s,p)\}$

(Eggstaff, Mazzuchi, & Sarkani, 2014)

A *calibration score* as defined above equal to 0.05 or higher **will suggest that the expert probabilities are statistically supported by the set of seed variables provided** (Cooke & Goosens, 2004). To determine the *information score* as depicted in Equation 6, we use the intrinsic range information already established, along with the established quantiles of (0.05, 0.50 and 0.95).

Equation 6 - Information Score Equation

$$inf(e) \approx -1 + \frac{1}{n}\sum_{i=1}^{n} ln\left(\frac{b_i - a_i}{(X_{i,5} - a_i)^{0.05}(X_{i,50} - X_{i,5})^{0.45}(X_{i,95} - X_{i,50})^{0.45}(b_i - X_{i,95})^{0.05}}\right)$$

(Eggstaff, Mazzuchi, & Sarkani, 2014)

To determine the global weight, expert calibration and relative information scores are multiplied together as depicted in Equation 7. The indicator function 1_α gets assigned a value of 0 or 1, depending on the cut-off level α determined during the analysis. If the expert calibration score is less than the cut-off level, then $1_\alpha = 0$, otherwise $1_\alpha = 1$.

Equation 7 - Global Weight Equation

$$w(e) = 1_\alpha \times cal(e) \times \inf(e)$$

The presence of a cut-off level α, is a requirement to maintain proper scoring rules within the SEJ method (Colson & Cooke, 2017). A cut-off level α is chosen during analysis to maximize the combined score of the resulting DM. Where the DM is defined as:

Equation 8 - Combined expert scores (DM)

$$DM_\alpha(i) = \frac{\sum_1^e w(e) \times s_{e,i}}{\sum_1^e w(e)}$$

(Colson & Cooke, 2017)

Optimizing the DM, often causes experts to be unweighted, even experts with good calibration scores. ***Such experts are not rejected***. The SEJ method has already captured those unweighted experts input in smaller subsets. The actual value of their input is

reflected during the robustness analysis for the "altered" DM with excluded experts (Colson & Cooke, 2017).

The data collected during the calibration and elicitation phases will be analyzed using a software tool developed by Cooke and the Technical University Delft. The software tool is called *Excalibur*, refer to Chapter 4. Next steps will prepare the *Spartan Team* for the dry-run and following process steps.

3.2.2.2 Dry-Run

During the dry-run phase two goals were set to be accomplished, (1) gave the *Spartan Team* a deeper understanding of the calibration phase coming up and (2) performance some optimization in the way the case structure and the target and query variables were presented to the *Spartan Team*. Overall the dry-run was a success as both objectives were accomplished, the feedback received was instrumental for optimizing the case structures, and more effectively focus on the uncertainty currently in the Project P1 and P2 target and query variables raw data while reducing survey ambiguity.

3.2.2.3 Calibration Survey

The purpose of the calibration phase is to obtain enough empirical data to support the hypothesis postulated in 3.2.2.1.5, for each of the participant being calibrated. The process for which the participant will get calibrated in this research study started with the calibration questions which are depicted in Appendix A. During the calibration process; each participant was given a time slot to attend to and the calibration survey. Each survey included the following:

- Ten (10) calibration questions related to the case structure, seed variables, and target questions identified by the student investigator.

- One (1) question asked to write the start and end time, to track the duration of the calibration survey.

To aid the participants answer the calibration questions the Risk Exposure (RE) tool is used, as depicted in Figure 11. This tool stands for a *response scale* and is one of the direct methods mentioned by (O'Hagan & et.al., 2006). A response scale provides a visual representation of the feasible range of responses. The response categories are based on the basic risk definition, which defines the risk as a triplet: a case (scenario), probability and impact. In the Project Management Institute (PMI) Lexicon of Project Management Terms, risk exposure is defined as an aggregate measure of the potential impact of all risks considered by the practitioner (Project Management Institute, 2017). In the context of this research, the participants are requested to review the RE to the case structure as defined initially for Projects P1 and P2.

Risk Exposure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	Minimal impact to the project					Minor impact to the project					Moderate impact to the project					Significal impact to the project					Critical impact to the project				

Figure 11 - Risk Exposure Tool (Response Scale)

Once all participants had completed the calibration survey the next step was to input all data collected into Excalibur for analysis.

3.2.2.3.1 Explaining an Excalibur Run

A dataset is created with the purpose of entering the data into Excalibur. Excalibur is a free software initially developed by Technical University Delft. The software helps the Classical Model analysis and determines other tests of importance. Depending on the user defined settings, which can vary depending on the options selected at the time of analysis.

These options are depicted in Figure 13. When the user executes an Excalibur RUN, it is calculating the calibration and information scores explained in section 3.2.2 for each of the experts in the research study. Also, it will calculate a synthetic expert named the decision maker (DM), and it will output other analysis such as a discrepancy and robustness tests for the dataset being analyzed. Both tests are explained in sections 3.2.4.1 and 3.2.4.2 respectively.

3.2.3 Elicitation Phase

The focus of the elicitation is to determine if the calibrated experts could indeed provide a better answer to the previous one by way of a structured method. Based on the target variables found in Table 11, the recently calibrated experts were ready to start the elicitation phase. All elicitation sessions were to be conducted individually to maintain neutrality among the participants. Figure 12 depicts a sample of the elicitation survey. Each survey asked the calibrated expert to answer the following:

- Four (4) elicitation questions related to the case structure and target variables and questions identified by the student investigator.
- Write down the start and end time, to track the duration of the elicitation survey.

Once again Excalibur is being used to input the calibrated expert assessment into the data set. For each expert and new elicitation variable, the assessment has to be entered

individually in Excalibur in the Assessment section as depicted in Figure 12. Important to note that the target variables for the elicitation did not have any realization values to enter.

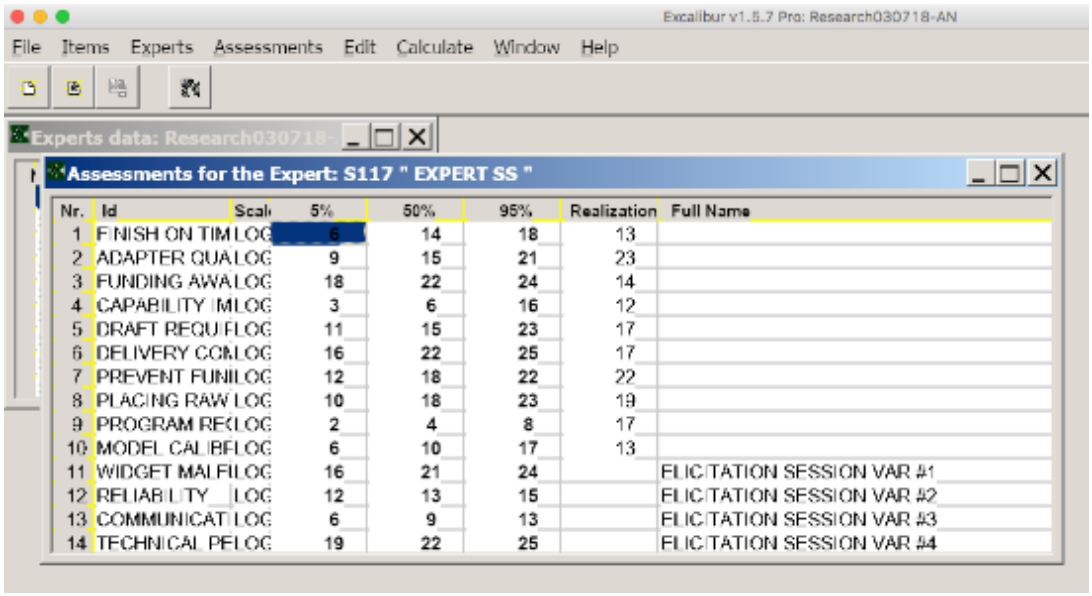


Figure 12 - Expert Assessment in Excalibur

The analysis is then run, once all the new variables from the new assessments are loaded into the dataset. The post analysis and relative information score are discussed in the next chapter.

3.2.4 Post-Elicitation

3.2.4.1 Discrepancy Test

A discrepancy analysis gets performed to determine how much all the calibrated experts agree or disagree with the synthetic expert (decision maker) (Cooke & Goossens, 1999). Using Excalibur and the current data set already collected, the discrepancy analysis is done during the actual Excalibur "run" this is depicted in Figure 13.

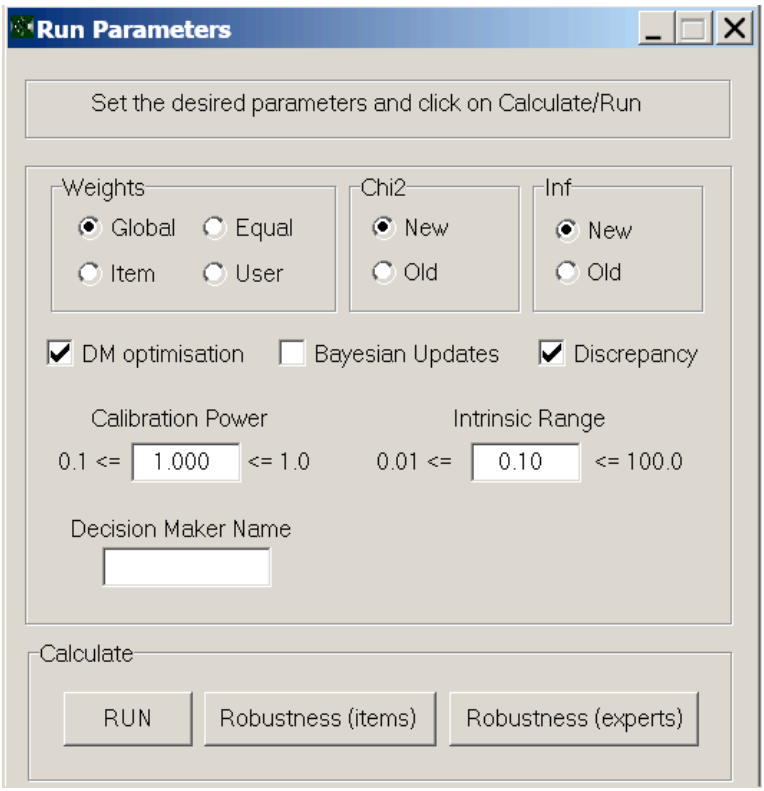


Figure 13 - Option GUI during Excalibur Run

3.2.4.2 Robustness Test

A robustness analysis tries to determine how strong an agreement there is between the expert decisions, the questions and the decision-maker (DM) (Cooke & Goossens, 1999). For robustness on the experts, this is carried out by removing one expert at a time and compared to the original DM. The best indicator of robustness is obtained by comparing the calibration and relative information of the "altered" DMs with respect to the original DM (Cooke & Jager, 1998).

3.3 Data Analysis Method (Evidence-Based)

The Bayesian inference method applicability is further explored to expand the data analysis efforts to improve probability estimates within the basic risk model depicted in Figure 9 when there is data or evidence available to make sound judgments in the

probability assessment. The available project data, which is provided by Project P4, will be used to drive this secondary goal, understand how Bayesian inferences can be applied using an optimal model. *Inferences are performed by sampling from a target distribution (i.e., a specially constructed Markov chain, based upon the inference problem) until convergence (to the posterior distribution) is achieved* (Kelly, Smith, Dezfuli, Vedros, & Galyean, 2009, p. 47). To support the inference the process flow map depicted in Figure 14 suggests how a proper target distribution model is selected based on the data type. When transferring this method to OpenBUGS, the modeling and the selection process will follow the recommendations in Figure 14, as well as a set of assumptions that will be described in the following sections.

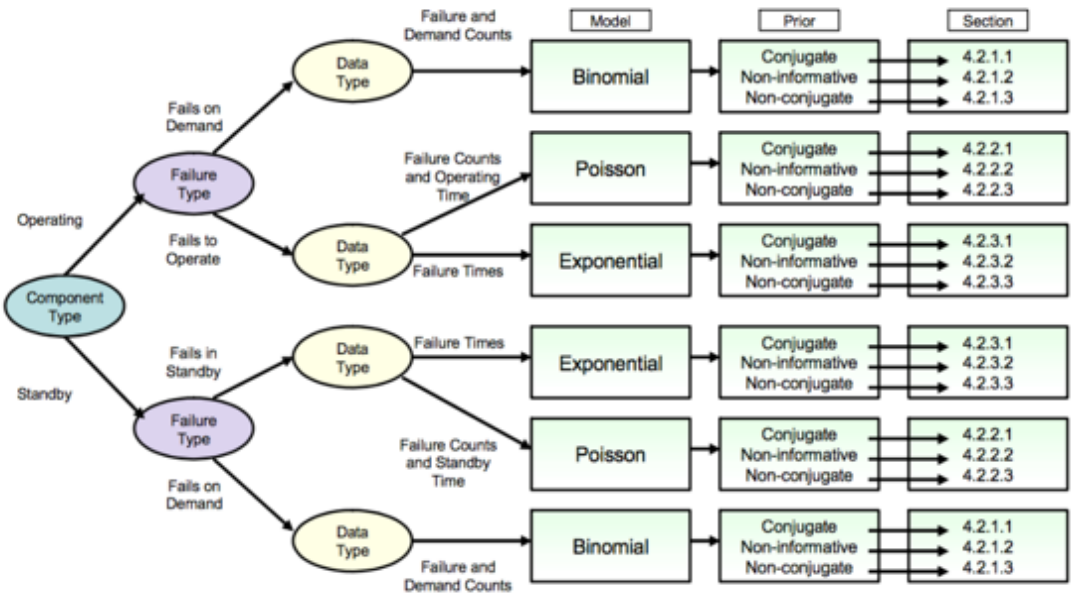


Figure 14 - Flow diagram for analysis method (Kelly and et al. 2009)

In addition to the suggested target distributions depicted in Figure 14, the method will explore the use of Lognormal and Weibull distributions as possible optimal models given the apparent nature of the dataset from Project P4.

3.3.1 Bayesian Inference (BI)

The Bayesian inference (BI) methodology is an evidence-based method that uses the Bayes theorem as the basis to determine any subjective probabilities and uses the same principles established in the Probabilistic Assessment Guidebook by (Kelly & Smith, 2011). The Bayes theorem is defined in Equation 9 below.

Equation 9 - Bayes Theorem

$$P(A|B) = P(A)\frac{P(B|A)}{P(B)}$$

The elements of the Bayes theorem are broken down and explained in Table 14. The following assumptions are critical in understanding the use of the Bayesian theorem in addressing the inference. The assumptions are:

- The application related to the use of the Bayesian inference will fit an aleatory model in the form of binomial, Poisson or exponential distributions and that these models have conjugate models.
- It is assumed that the prior distribution is impacted by the lack of data, to account for this impact a Jeffrey's prior appropriate for the aleatory model is introduced.

Table 14 - Bayes Theorem Elements

Element	Description
P(A|B)	Posterior distribution: Conditional upon the data available in B and how it relates to the hypothesis A
P(A)	Prior distribution: From the knowledge of the hypothesis A and independent of data B

P(B|A)	Posterior Distribution: Conditional upon the data available in B
P(B)	Marginal Distribution of data B

Based on the Bayes theorem elements and the assumptions, the following are the aleatory models that make the foundation of the method going forward. The underlining assumption of this approach and the basis of using BI is that all the data and information is non-informative. This assumption will allow the BI to use Jeffrey's prior for all the conjugate priors defined for the three common distributions.

3.3.1.1 Binomial and Beta Distributions

The binomial distribution and its conjugate prior and beta distribution are defined in Equation 10 and Equation 11.

Equation 10 - Binomial Distribution

$$\Pr(x|p) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \dots, n$$

Where *p*, is the probability of the event, and *n* are the number of trials.

Equation 11 - Beta Distribution

$$f(y, \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} y^{\alpha-1}(1-y)^{\beta-1}, \quad 0 \le y \le 1$$

Where α and β are the shape parameters in Equation 11. The shape parameters that will define the Jeffrey's prior for the non-informative data or information provided are showed in Equation 12:

Equation 12 - Jeffrey's Prior Parameters for Beta Distribution

$$\alpha = 0.5$$
$$\beta = 0.5$$

3.3.1.2 Poisson and Gamma Distributions

The Poisson distribution and its conjugate prior a gamma distribution are defined in Equation 13 and

The rate in the Poisson process is defined by the λ variable based on the number of events *x*.

Equation 14.

Equation 13 - Poisson Distribution

$$\Pr(x) = \frac{\lambda t^x}{x!} e^{-\lambda}, \quad for\ x = 0,1, \ldots$$

The rate in the Poisson process is defined by the λ variable based on the number of events *x*.

Equation 14 - Gamma Distribution

$$f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-t\beta}$$

Where α and β are the shape parameters in
The rate in the Poisson process is defined by the λ variable based on the number of events *x*.

Equation 14 and *t* is the time domain variable. The shape parameters that will define the Jeffrey's prior for the non-informative data or information supplied are showed in Equation 15:

Equation 15 - Jeffrey's Parameters for a Gamma Distribution

$$\alpha = 0.5$$
$$\beta = 0 \; or \; 0.0001 \; for \; numerical \; analysis$$

Using a beta parameter of zero will result in an improper distribution, but this research study is using numerical analysis methods, more specifically OpenBUGS to solve the models, resulting in proper distributions.

3.3.1.3 Exponential and Gamma Distributions

The exponential distribution and its conjugate prior a gamma distribution are defined in Equation 16 below and

The rate in the Poisson process is defined by the λ variable based on the number of events *x*.

Equation 14 in the prior section.

Equation 16 - Exponential Distribution

$$f(t) = \lambda e^{-\lambda t}$$

The λ variable is the rate in the exponential process within a *t* time domain. The shape parameters that will define the Jeffrey's prior for the non-informative data or information provided are showed in Equation 17:

Equation 17 - Jeffrey's Parameters for a Gamma Distribution (Exponential Case)

$$\alpha = 0 \; or \; 0.0001 \; for \; numerical \; analysis$$
$$\beta = 0 \; or \; 0.0001 \; for \; numerical \; analysis$$

As with the Poisson distribution, the resulting prior is an improper distribution, but it will result in a proper distribution when resolved numerically.

3.3.1.4 Weibull and Lognormal Distributions

For a random variable, t, that has a Weibull distribution the probability density function (PDF) is defined in Equation 18, (Kelly & Smith, 2011).

Equation 18 - Weibull Distribution PDF

$$f(t) = \frac{\beta}{\alpha}\left(\frac{t}{\alpha}\right)^{\beta-1} e^{-\left(\frac{t}{\alpha}\right)^{\beta}}$$

$$\alpha\ parameter = the\ scale\ of\ the\ distribution$$
$$\beta\ paramter = shape\ of\ the\ distribution$$
$$t = time\ domain\ variable$$

The T distribution is said to be lognormally distributed when $T = \ln(t)$. The PDF for the lognormal distribution is defined in Equation 19.

Equation 19 - Lognormal Distribution

$$\mathrm{f(t)} = \frac{1}{\sigma t\sqrt{2\pi}} e^{\left(-\frac{1}{2\sigma^2}(\ln(t)-\mu)^2\right)}$$

The mean and variance of the lognormal distribution are denoted by the variables μ, and σ^2 respectively.

3.3.1.5 Model Checking with Summary Statistics and Bayesian Information

To determine if the *predicted values* of the BI model align to the *observed values*, in this case, given by Project P4 dataset, the frequentists approach to model checking, in general, involves comparing the observed value of a test statistic to a p-value. The Bayesian Chi-square statistic for observed values is defined as follows:

Equation 20 - Chi-square test statistic for observed values

$$\chi^2_{obs} = \sum_i \frac{\left(x_{obs,i} - \mu_i\right)^2}{\sigma_i^2}$$

(Kelly & Smith, 2011)

In this equation $x_{obs,i}$ represents the observed values, μ is the mean of the observed values and σ^2 the variance. A similar equation is defined for the predicted values as shown below:

Equation 21 - Chi-square test statistic for predicted values

$$\chi^2_{pred} = \sum_i \frac{\left(x_{pred,i} - \mu_i\right)^2}{\sigma_i^2}$$

(Kelly & Smith, 2011)

The analogous Bayesian approach to this method involves using the summary statistics from OpenBUGS which are derived from the posterior distribution (Kelly & Smith, 2011) and an analysis method called Cramer-von Mises (CVM) statistic. The CVM compares the empirical cumulative distribution function for the ranked observed and replicated durations and calculates a p-value analogous, or *Bayes factor (BF)*, to the frequentists p-value (Kelly & Smith, 2011). In this case the BF is defined as the probability that the predicted values for a chi-square test statistic will be equal or greater than those of an observed one, as shown below:

Equation 22 - Bayes Factor Definition

$$\Pr\left(\chi^2_{pred} \geq \chi^2_{obs}\right)$$

(Kelly & Smith, 2011)

The Bayesian Information Criterion (BIC) is recommended for selecting the optimal probability distribution among non-hierarchical models (Kelly & Smith, 2011), which is

the case in this research study. The BIC is defined in Equation 23 and implemented using OpenBUGS.

Equation 23 - Bayesian Information Criterion (BIC)

$$\text{BIC} = -2 \times \log(likelihood) + k \times log(N)$$

In this equation, *likelihood* represents the Bayesian likelihood function, *k* is the number of distributions being compared and *N* is the number of data points (Kelly & Smith, 2011). To assess the resultant value BIC provides during the model checking the model with the smallest BIC is preferred (Kelly & Smith, 2011).

3.3.1.6 OpenBUGS Modeling Scripts

The modeling tool selected to aid in the numerical solution of the BI is OpenBUGS. This tool uses a scripting language and uses the Monte Carlo Markov Chain (MCMC) as the underlining algorithm. The model distribution scripts used for the optimization are all documented in Appendix E.

3.3.1.7 Hardware Codes

The model predictions will be run in four unique hardware failure schemes, given by the donating project P4. The assigned hardware codes are depicted in Table 15.

Table 15 - Hardware Codes from Project P4

Hardware Code	
Widget A	Project P4 Hardware #1
Widget B	Project P4 Hardware #2
Widget C	Project P4 Hardware #3
Widget D	Project P4 Hardware #4

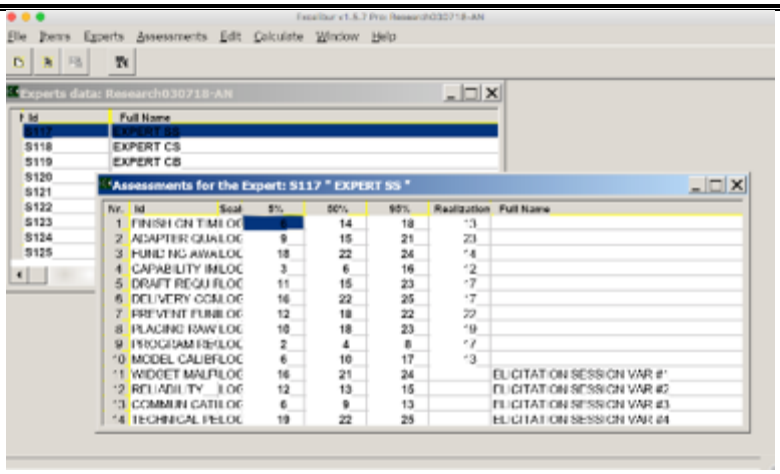
Chapter 4 Results

This chapter presents the results for both the calibration and elicitation of the SEJ method as well as the results for the BI method. The results are divided into four distinct sections, were Section 4.1 shows the expert assessment and data collected for the SEJ method. Section 4.2 shows the expert performance results, which includes the calibration and information scores and the PWDM results, as well as the results from the robustness and discrepancy tests. Section 4.3 will show the results for the BI method and its probability distribution predictions. Finally, section 4.4 will show the results for the hypothesis tests.

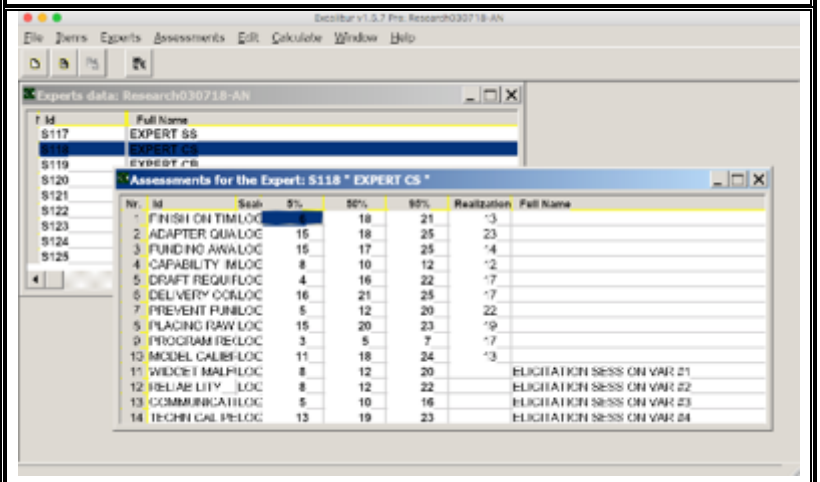
4.1 Expert Assessments and Data Collection

This section holds the assessment of all the experts S117 to S132 after completing all elicitation phases. The data collected from the experts was tabulated and entered into Excalibur for data analysis. Excalibur supplied the initial processing of the data collection and output the calibration, information scores, and the PWDM as well as additional data for post-processing in Statistical Package for the Social Sciences (SPSS) software. Table 16 below is the compilation of all the data collected during the elicitation.

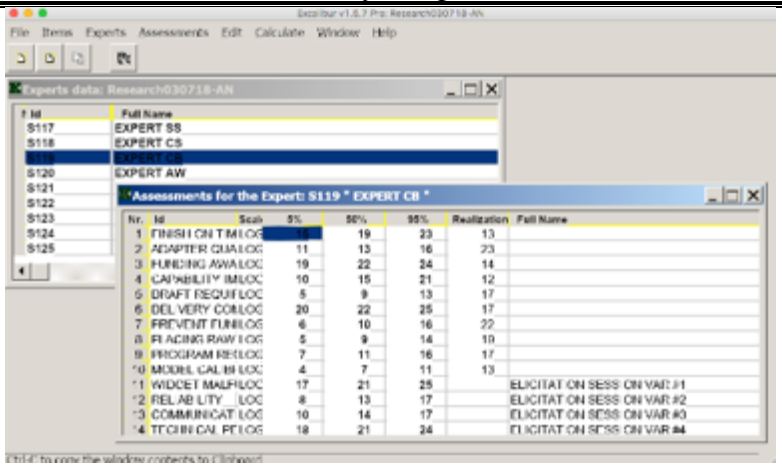
Table 16 - Complete SEJ Assessment for Experts S117 to S132



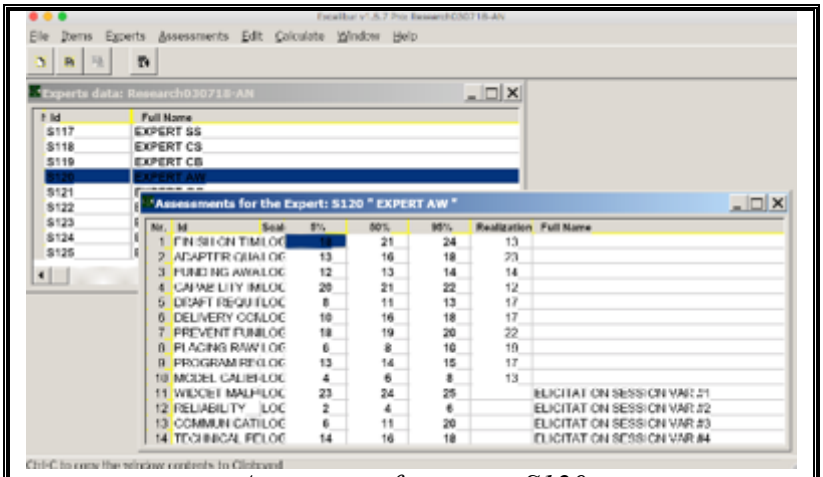
Assessment for expert S117



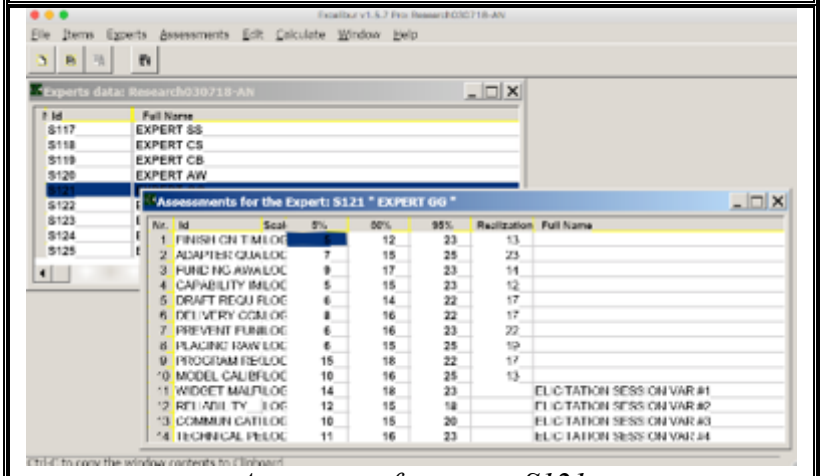
Assessment for expert S118



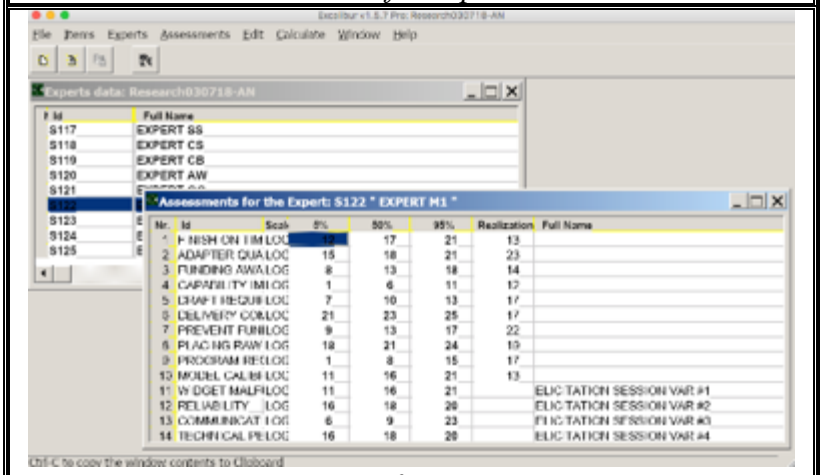
Assessment for expert S119



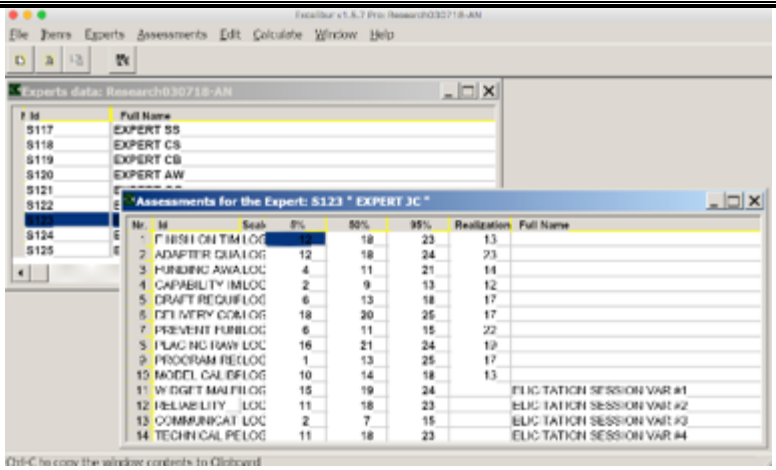
Assessment for expert S120



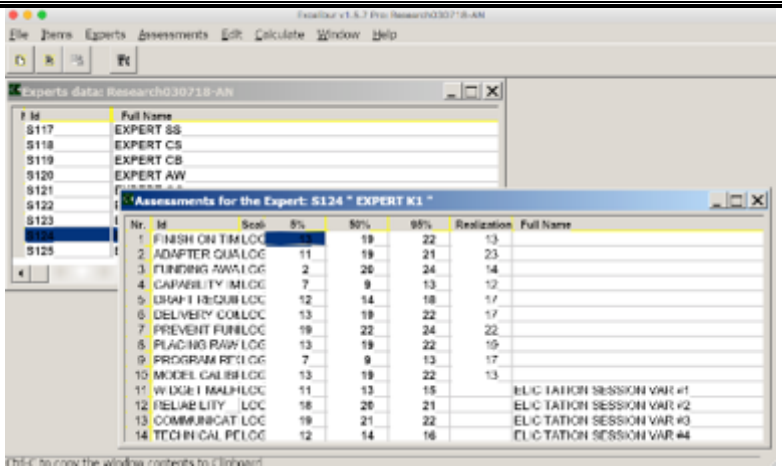
Assessment for expert S121



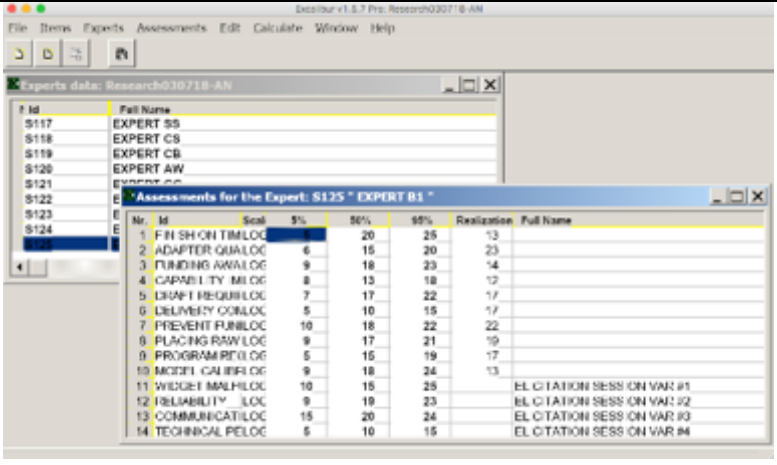
Assessment for expert S122



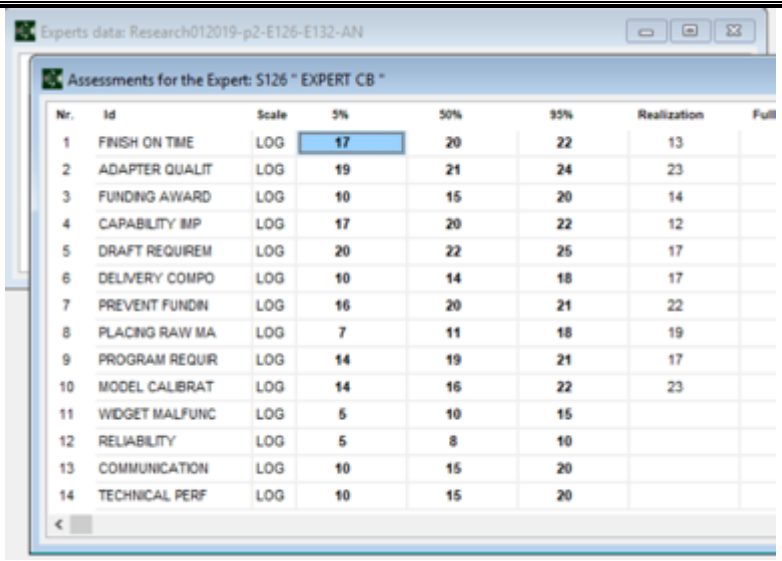
Assessment for expert S123



Assessment for expert S124

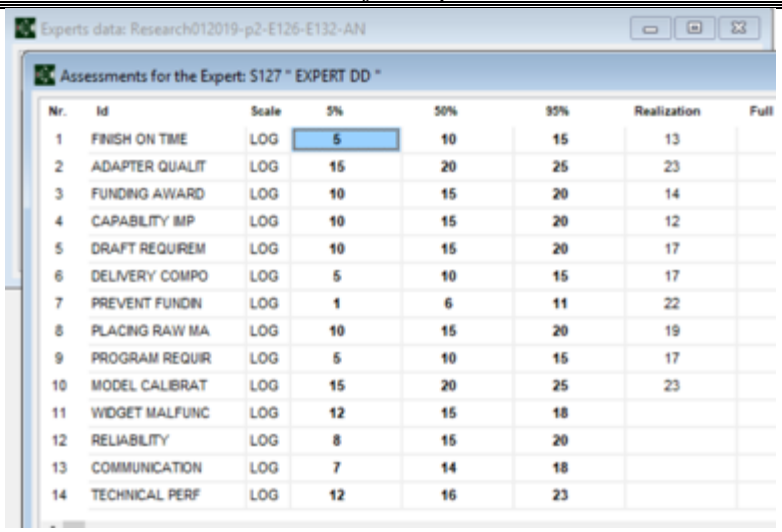


Assessment for expert S125



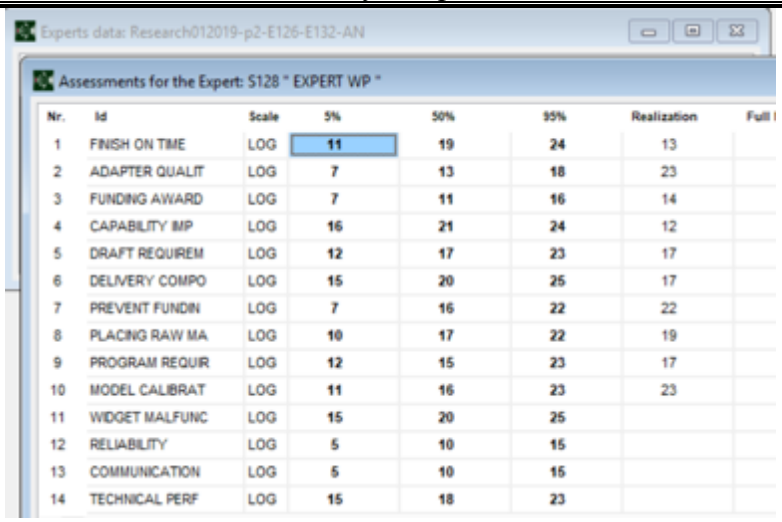
Nr.	Id	Scale	5%	50%	95%	Realization	Full
1	FINISH ON TIME	LOG	17	20	22	13	
2	ADAPTER QUALIT	LOG	19	21	24	23	
3	FUNDING AWARD	LOG	10	15	20	14	
4	CAPABILITY IMP	LOG	17	20	22	12	
5	DRAFT REQUIREM	LOG	20	22	25	17	
6	DELIVERY COMPO	LOG	10	14	18	17	
7	PREVENT FUNDIN	LOG	16	20	21	22	
8	PLACING RAW MA	LOG	7	11	18	19	
9	PROGRAM REQUIR	LOG	14	19	21	17	
10	MODEL CALIBRAT	LOG	14	16	22	23	
11	WIDGET MALFUNC	LOG	5	10	15		
12	RELIABILITY	LOG	5	8	10		
13	COMMUNICATION	LOG	10	15	20		
14	TECHNICAL PERF	LOG	10	15	20		

Assessment for expert S126



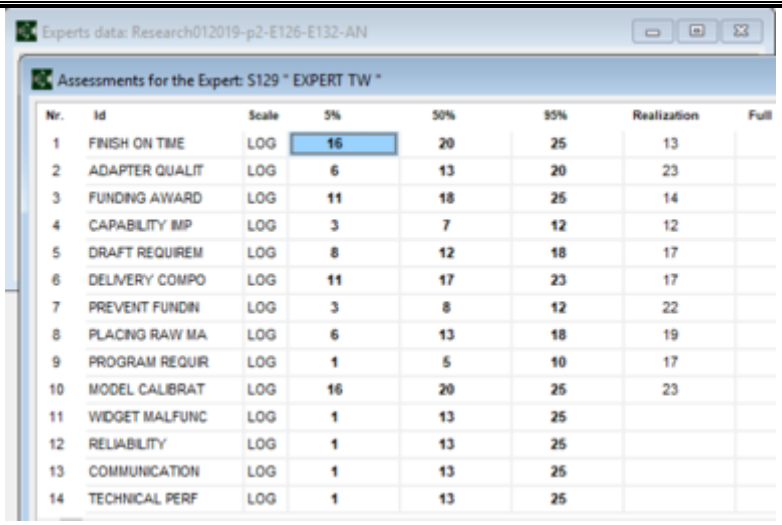
Nr.	Id	Scale	5%	50%	95%	Realization	Full
1	FINISH ON TIME	LOG	5	10	15	13	
2	ADAPTER QUALIT	LOG	15	20	25	23	
3	FUNDING AWARD	LOG	10	15	20	14	
4	CAPABILITY IMP	LOG	10	15	20	12	
5	DRAFT REQUIREM	LOG	10	15	20	17	
6	DELIVERY COMPO	LOG	5	10	15	17	
7	PREVENT FUNDIN	LOG	1	6	11	22	
8	PLACING RAW MA	LOG	10	15	20	19	
9	PROGRAM REQUIR	LOG	5	10	15	17	
10	MODEL CALIBRAT	LOG	15	20	25	23	
11	WIDGET MALFUNC	LOG	12	15	18		
12	RELIABILITY	LOG	8	15	20		
13	COMMUNICATION	LOG	7	14	18		
14	TECHNICAL PERF	LOG	12	16	23		

Assessment for expert S127



Nr.	Id	Scale	5%	50%	95%	Realization	Full
1	FINISH ON TIME	LOG	11	19	24	13	
2	ADAPTER QUALIT	LOG	7	13	18	23	
3	FUNDING AWARD	LOG	7	11	16	14	
4	CAPABILITY IMP	LOG	16	21	24	12	
5	DRAFT REQUIREM	LOG	12	17	23	17	
6	DELIVERY COMPO	LOG	15	20	25	17	
7	PREVENT FUNDIN	LOG	7	16	22	22	
8	PLACING RAW MA	LOG	10	17	22	19	
9	PROGRAM REQUIR	LOG	12	15	23	17	
10	MODEL CALIBRAT	LOG	11	16	23	23	
11	WIDGET MALFUNC	LOG	15	20	25		
12	RELIABILITY	LOG	5	10	15		
13	COMMUNICATION	LOG	5	10	15		
14	TECHNICAL PERF	LOG	15	18	23		

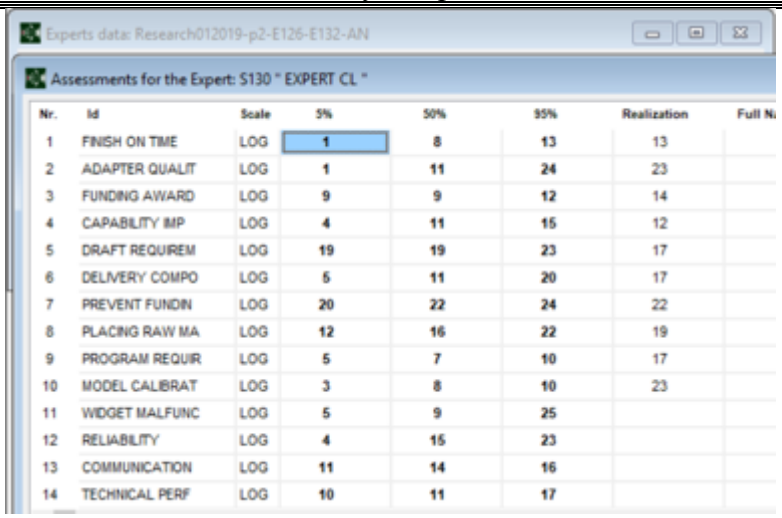
Assessment for expert S128


Experts data: Research012019-p2-E126-E132-AN

Assessments for the Expert: S129 " EXPERT TW "

Nr.	Id	Scale	5%	50%	95%	Realization	Full
1	FINISH ON TIME	LOG	16	20	25	13	
2	ADAPTER QUALIT	LOG	6	13	20	23	
3	FUNDING AWARD	LOG	11	18	25	14	
4	CAPABILITY IMP	LOG	3	7	12	12	
5	DRAFT REQUIREM	LOG	8	12	18	17	
6	DELIVERY COMPO	LOG	11	17	23	17	
7	PREVENT FUNDIN	LOG	3	8	12	22	
8	PLACING RAW MA	LOG	6	13	18	19	
9	PROGRAM REQUIR	LOG	1	5	10	17	
10	MODEL CALIBRAT	LOG	16	20	25	23	
11	WIDGET MALFUNC	LOG	1	13	25		
12	RELIABILITY	LOG	1	13	25		
13	COMMUNICATION	LOG	1	13	25		
14	TECHNICAL PERF	LOG	1	13	25		

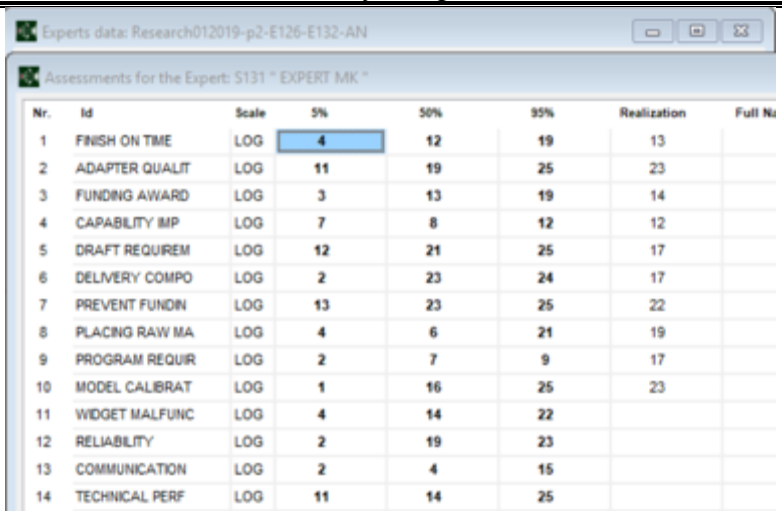
Assessment for expert S129


Experts data: Research012019-p2-E126-E132-AN

Assessments for the Expert: S130 " EXPERT CL "

Nr.	Id	Scale	5%	50%	95%	Realization	Full Nu
1	FINISH ON TIME	LOG	1	8	13	13	
2	ADAPTER QUALIT	LOG	1	11	24	23	
3	FUNDING AWARD	LOG	9	9	12	14	
4	CAPABILITY IMP	LOG	4	11	15	12	
5	DRAFT REQUIREM	LOG	19	19	23	17	
6	DELIVERY COMPO	LOG	5	11	20	17	
7	PREVENT FUNDIN	LOG	20	22	24	22	
8	PLACING RAW MA	LOG	12	16	22	19	
9	PROGRAM REQUIR	LOG	5	7	10	17	
10	MODEL CALIBRAT	LOG	3	8	10	23	
11	WIDGET MALFUNC	LOG	5	9	25		
12	RELIABILITY	LOG	4	15	23		
13	COMMUNICATION	LOG	11	14	16		
14	TECHNICAL PERF	LOG	10	11	17		

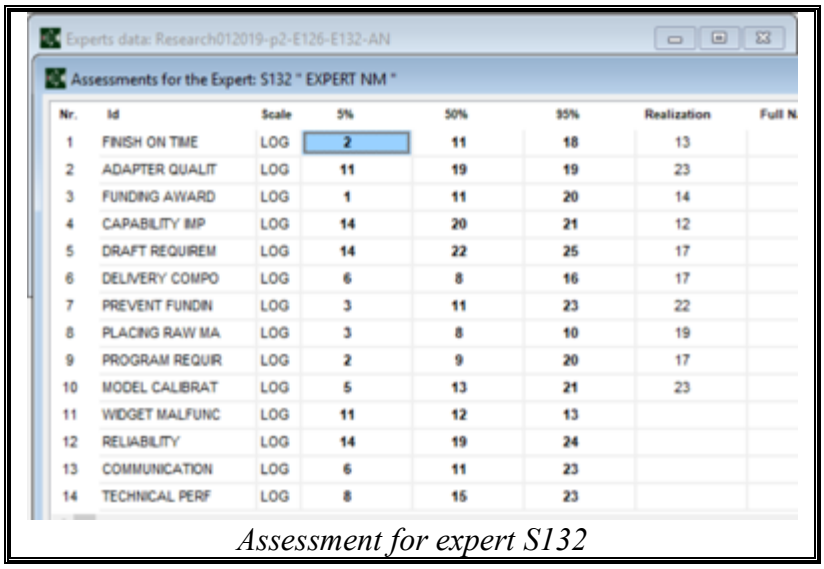
Assessment for expert S130


Experts data: Research012019-p2-E126-E132-AN

Assessments for the Expert: S131 " EXPERT MK "

Nr.	Id	Scale	5%	50%	95%	Realization	Full Nu
1	FINISH ON TIME	LOG	4	12	19	13	
2	ADAPTER QUALIT	LOG	11	19	25	23	
3	FUNDING AWARD	LOG	3	13	19	14	
4	CAPABILITY IMP	LOG	7	8	12	12	
5	DRAFT REQUIREM	LOG	12	21	25	17	
6	DELIVERY COMPO	LOG	2	23	24	17	
7	PREVENT FUNDIN	LOG	13	23	25	22	
8	PLACING RAW MA	LOG	4	6	21	19	
9	PROGRAM REQUIR	LOG	2	7	9	17	
10	MODEL CALIBRAT	LOG	1	16	25	23	
11	WIDGET MALFUNC	LOG	4	14	22		
12	RELIABILITY	LOG	2	19	23		
13	COMMUNICATION	LOG	2	4	15		
14	TECHNICAL PERF	LOG	11	14	25		

Assessment for expert S131



Nr.	Id	Scale	5%	50%	95%	Realization	Full N
1	FINISH ON TIME	LOG	2	11	18	13	
2	ADAPTER QUALIT	LOG	11	19	19	23	
3	FUNDING AWARD	LOG	1	11	20	14	
4	CAPABILITY IMP	LOG	14	20	21	12	
5	DRAFT REQUIREM	LOG	14	22	25	17	
6	DELIVERY COMPO	LOG	6	8	16	17	
7	PREVENT FUNDIN	LOG	3	11	23	22	
8	PLACING RAW MA	LOG	3	8	10	19	
9	PROGRAM REQUIR	LOG	2	9	20	17	
10	MODEL CALIBRAT	LOG	5	13	21	23	
11	WIDGET MALFUNC	LOG	11	12	13		
12	RELIABILITY	LOG	14	19	24		
13	COMMUNICATION	LOG	6	11	23		
14	TECHNICAL PERF	LOG	8	15	23		

Assessment for expert S132

4.2 Expert Performance

Once the data assessments as depicted in Table 16 were entered in Excalibur, the process to determine the calibration and information scores followed. It is essential to recall Figure 13, which depicts the options available for the standard Chi-square goodness of fit test analysis which determines the calibration and information scores. The analysis uses the following options:

- Global performance weight
- Calibration power of 1.0

These settings will determine the calibration and relative information score of synthetic experts, the PWDM. Once the run parameters are selected as depicted in Figure 16, the analysis can be run in Excalibur. An example of what makes a good a well-calibrated expert is provided by (Ryan & et al., 2012), as shown in Figure 15 which depicts an "*Expert 3*" as well-calibrated.

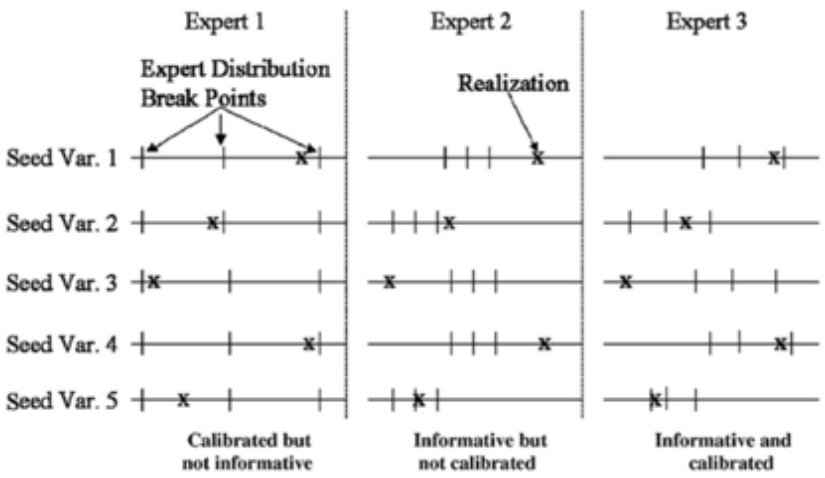


Figure 15 - Example of a Well-Calibrated Expert

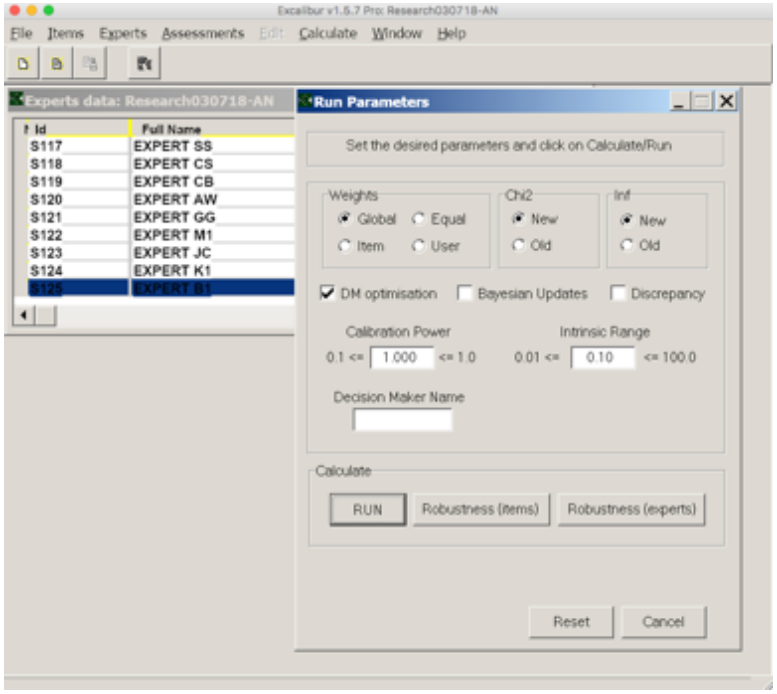


Figure 16 - Run parameters used for Excalibur

The resulting expert performance given the realization of the seed variables is shown in Table 17. Experts S117 to S132 are shown in the 2nd column. The next column shows the calibration score as calculated by Equation 5. The 4th column shows the expert

relative information score as calculated by Equation 6. The last two columns show the resultant unnormalized, as determined by Equation 7 and normalized weights which should add to one as a result of normalizing the expert and PWDM unnormalized weights.

Table 17 - Summary of Experts Performance (Seed Variables)

N	Experts	Calibration Score	Mean Information Score	Unnormalized Weight	Normalized Weight
1	S117	0.1917	1.0890	0.0000	0.0000
2	S118	0.3136	1.2300	0.0000	0.0000
3	S119	3.50E-08	1.3150	0.0000	0.0000
4	S120	4.94E-07	2.0200	0.0000	0.0000
5	S121	0.4735	0.8202	0.0000	0.0000
6	S122	0.0003	1.2720	0.0000	0.0000
7	S123	0.7069	1.0170	0.0000	0.0000
8	S124	0.0750	1.4550	0.0000	0.0000
9	S125	0.2441	0.9382	0.0000	0.0000
10	S126	0.0013	1.7520	0.0000	0.0000
11	S127	0.0471	1.0150	0.0000	0.0000
12	S128	0.8283	1.1790	0.9769	0.5000
13	S129	0.0013	0.9278	0.0000	0.0000
14	S130	0.0171	1.1240	0.0000	0.0000
15	S131	0.6827	0.9357	0.0000	0.0000
16	S132	0.0471	0.8698	0.0000	0.0000
17	PWDM	0.8283	1.1790	0.9769	0.5000

To provide a more in-depth perspective on the results from Table 17 above, a summary of the expert performance on seed variables, each expert's results (S117 to S132) are presented graphically in a range graph. The graph depicts each seed variable used during the calibration phase. Each graph's range is denoted by the intrinsic range defined in the RE tool. The PWDM's uncertainty range is shown in the second position from the

top, the first being the realization for the corresponding seed variable. Recall, Table 10 for the list of seed variables. Each line stands for the 5th, 50th, and 95th quantile estimate range given by experts S117 to S132. Finally, the x-axis is the elicited RE range for each target variable. The remaining range of graphs for the seed variables are shown in Appendix F: Range Graphs.

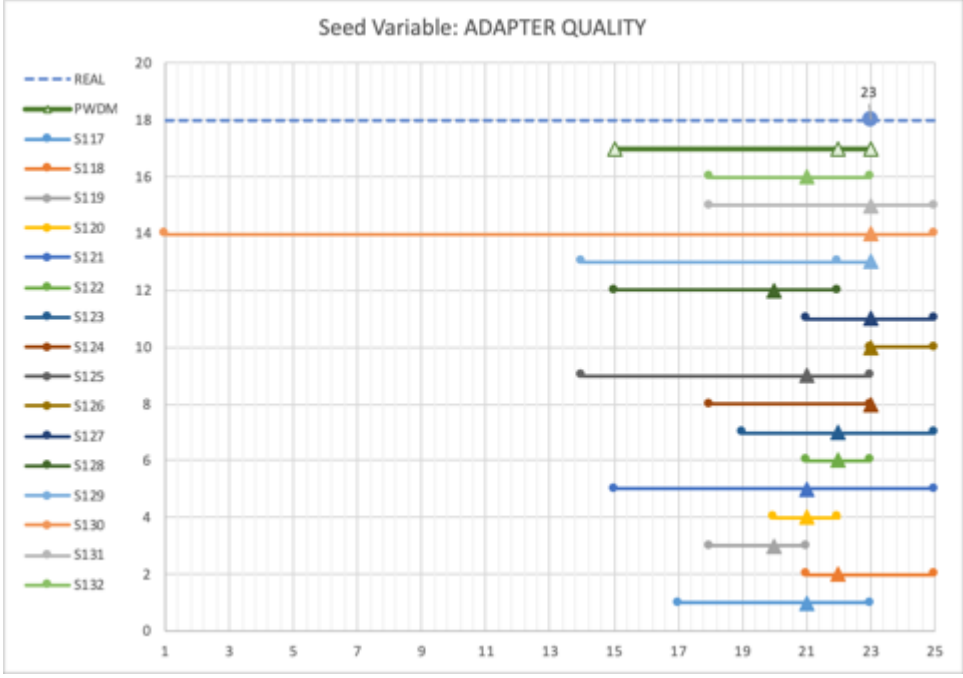


Figure 17 - Calibration Results on Adapter Quality

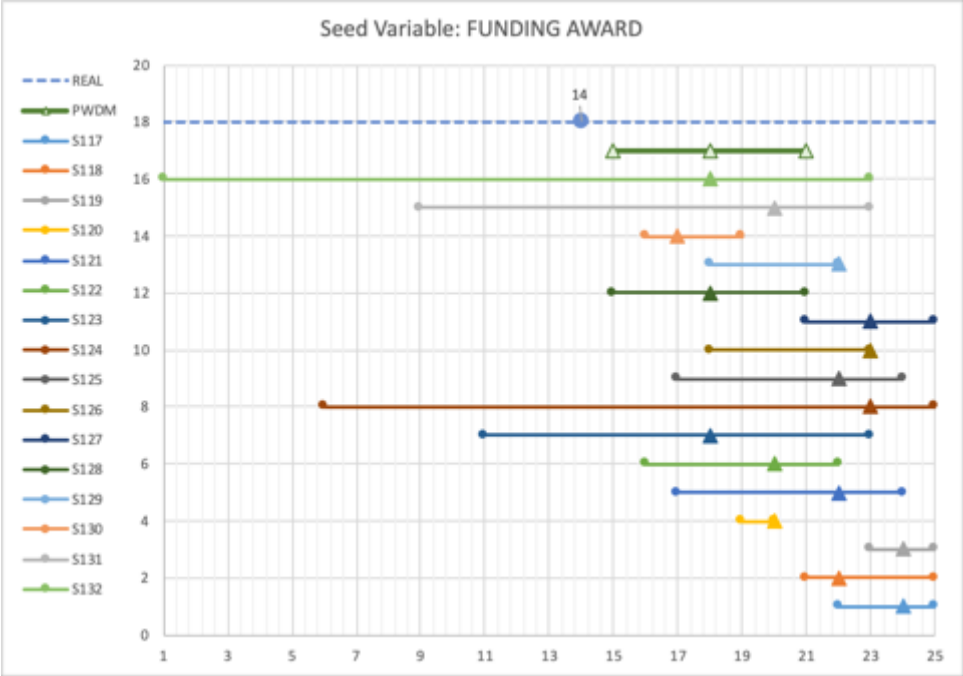


Figure 18 - Calibration Results on Funding Award

The expert performance on the target variables is depicted in Table 18. Experts S117 to S132 are shown in the 2nd column. The next column shows the calibration score as calculated by Equation 5. The 4th column shows the expert relative information score as calculated by Equation 6. The last two columns show the resultant unnormalized, as determined by Equation 7 and normalized weights which should add to one as a result of normalizing the expert and PWDM unnormalized weights.

Table 18 - Summary of Experts Performance (Target Variables)

N	Experts	Calibration Score	Mean Information Score	Unnormalized Weight	Normalized Weight
1	S117	2.54E-05	1.4800	0.0000	0.0000
2	S118	0.0220	0.7149	0.0000	0.0000

3	S119	2.54E-05	1.2930	0.0000	0.0000
4	S120	0.0018	1.3960	0.0000	0.0000
5	S121	2.54E-05	1.0900	0.0000	0.0000
6	S122	0.0018	1.3030	0.0000	0.0000
7	S123	0.0018	0.7853	0.0000	0.0000
8	S124	0.0018	1.9840	0.0000	0.0000
9	S125	0.0018	0.8441	0.0000	0.0000
10	S126	0.0765	0.8273	0.0000	0.0000
11	S127	2.54E-05	0.9513	0.0000	0.0000
12	S128	0.0220	0.8982	0.0000	0.0000
13	S129	0.0018	0.8738	0.0000	0.0000
14	S130	0.0765	0.8397	0.0000	0.0000
15	S131	0.4817	0.4213	0.2029	0.5000
16	S132	0.0018	1.0540	0.0000	0.0000
17	PWDM	0.4817	0.4213	0.2029	0.5000

The graphs depict each target variable used during the elicitation phase. Each graph's range is denoted by the intrinsic range defined in the RE tool. The PWDM's uncertainty range is shown in the second position from the top, the first being the realization for the corresponding target variable. Recall, Table 11 for the list of target variables. Each line represents the 5^{th}, 50^{th}, and 95^{th} quantile estimate range given by experts S117 to S132.

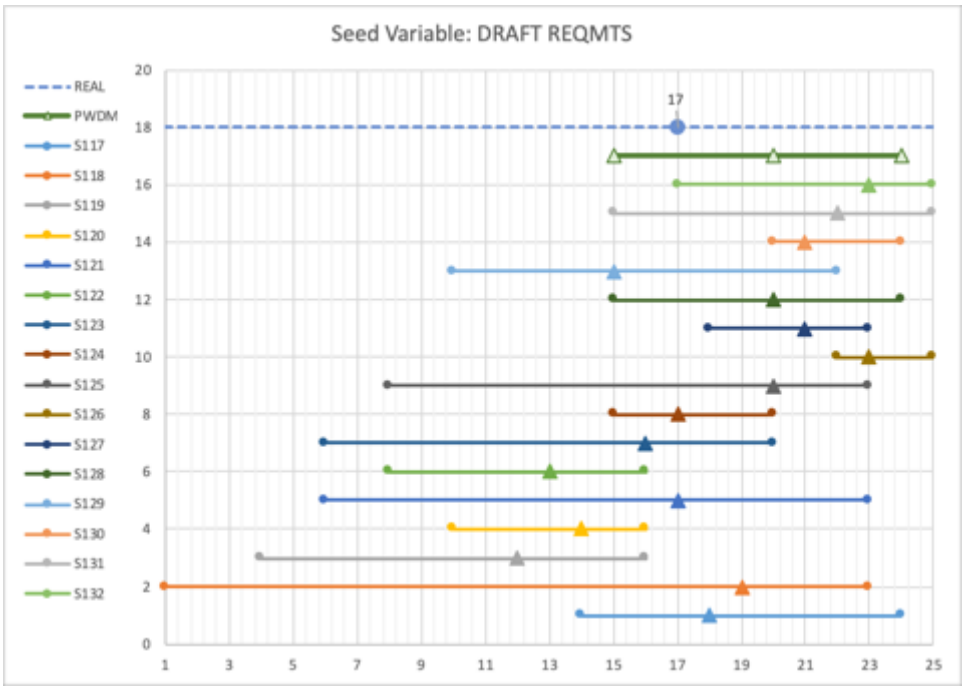


Figure 19 - Calibration Results on Draft RMF Requirement

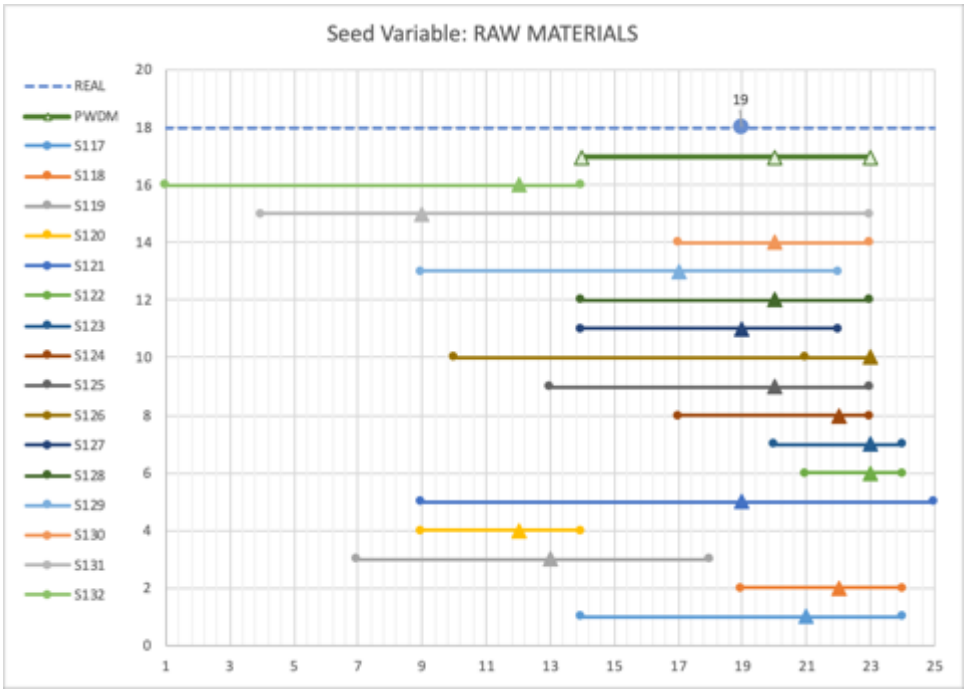


Figure 20 - Calibration Results on Placing Raw Materials Orders

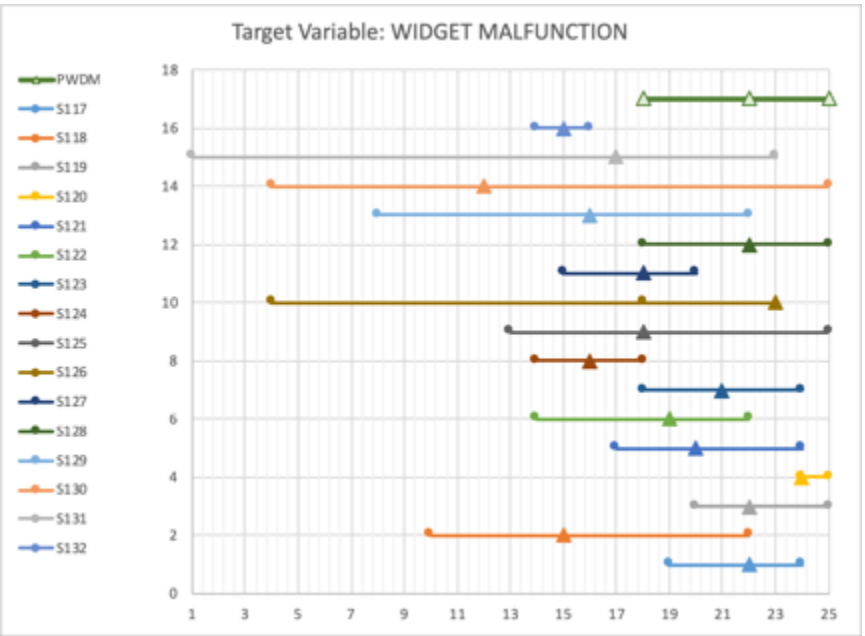


Figure 21 - Elicitation Results on Widget Malfunction

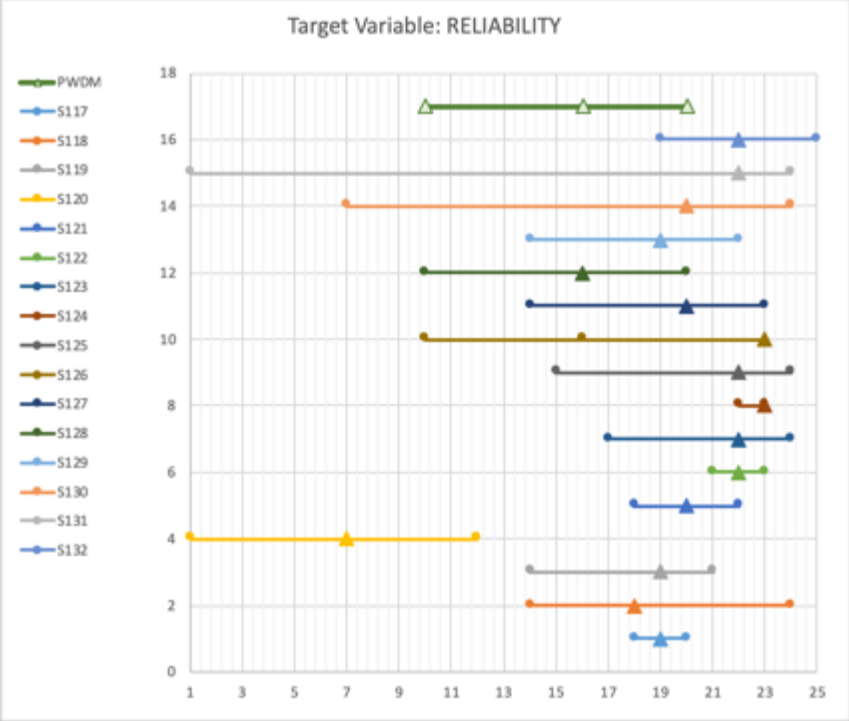


Figure 22 - Elicitation Results on Reliability

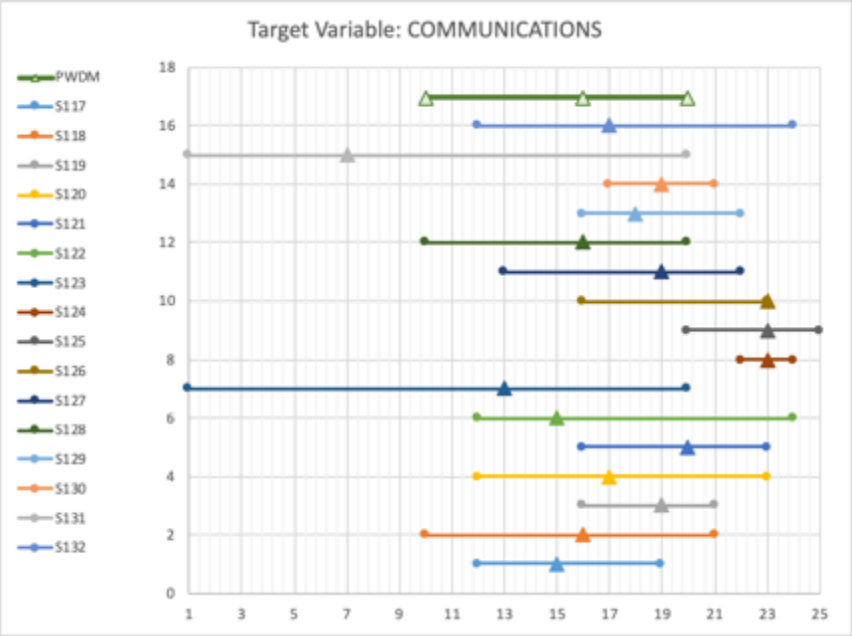


Figure 23 - Elicitation Results on Communication

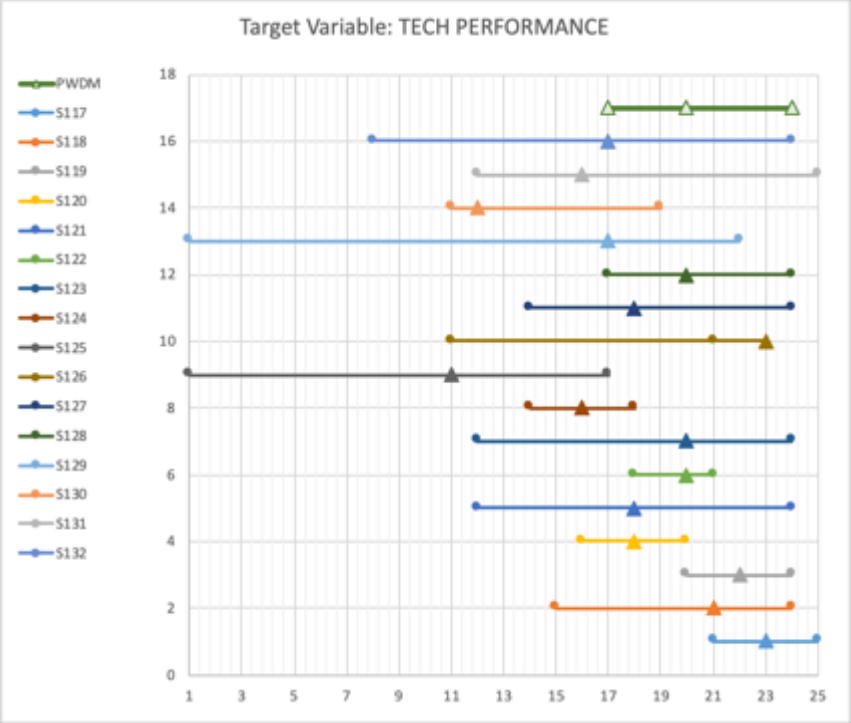


Figure 24 - Elicitation Results on Technical

4.2.1 Expert's Robustness Results

A robustness analysis is required to determine how individual experts are impacting the "Original" PWDM measure, reference section 3.2.4.2. This new measure the "Altered" PWDM is obtained by removing one expert at a time from the process and recalculating the "Altered" PWDM. The results are depicted in Table 19 below. The second and third columns show the results from the expert-robustness, but the third column measures how the "Altered" PWDM compares to the "Original" PWDM.

Table 19 – Expert Robustness Results

	Altered PWDM		Original PWDM
Excluded Expert	Calibration Score	Inf. Score	Inf. Score
S117	0.8283	1.1790	1.0890
S118	0.8283	1.1680	1.2300
S119	0.8283	1.1790	1.3150
S120	0.8283	1.1790	2.0200
S121	0.8283	1.1780	0.8202
S122	0.8283	1.1570	1.2720
S123	0.8283	1.1780	1.0170
S124	0.8283	1.1790	1.4550
S125	0.8283	1.1790	0.9382
S126	0.8283	1.1790	1.7520
S127	0.8283	1.1430	1.0150
S128	0.7069	1.0160	1.1790
S129	0.8283	1.1790	0.9278
S130	0.8283	1.0880	1.1240
S131	0.8283	1.1010	0.9357
S132	0.8283	1.1450	0.8698
None	0.8283	1.1790	1.1790

Similarly, to measure the robustness of the seed variables against the "Original" PWDM measure, the new "Altered" PWDM is obtained by removing one seed variable at a time from the process. The results are depicted in Table 20 below. The second and third columns show the results from the seed-robustness, but the third column measures how the "Altered" PWDM compares to the "Original" PWDM.

Table 20 – Seed Variables Robustness Results

	Altered PWDM		Original PWDM
Seed Variable Excluded	Calibration Score	Inf. Score	Inf. Score
1	0.7305	1.1700	0.0000
2	0.7305	0.9860	1.0360
3	0.7059	0.8024	0.4564
4	0.7305	1.0650	0.8216
5	0.7059	0.8303	0.4812
6	0.7305	1.1600	0.0000
7	0.5710	0.8372	0.8142
8	0.7305	1.2160	0.0000
9	0.5925	0.7567	0.6895
10	0.7305	1.1740	0.0000

4.2.2 Expert's Discrepancy Results

The discrepancy analysis aims to identify which of the expert's uncertainty assessments differ the most (Cooke & Goosens, 2004). Reference section 3.2.4.1. The results of the discrepancy analysis are depicted in Table 21 shown below.

Table 21 - Discrepancy Analysis Results

N	Experts	Calibration Score	Mean Information Score	Relative Discrepancy Ratio
1	S117	0.1917	1.0890	1.4600
2	S118	0.3136	1.2300	1.5020
3	S119	3.50E-08	1.3150	1.8060
4	S120	4.94E-07	2.0200	1.7710
5	S121	0.4735	0.8202	0.8422
6	S122	0.0003	1.2720	1.3240
7	S123	0.7069	1.0170	1.0620
8	S124	0.0750	1.4550	1.5390
9	S125	0.2441	0.9382	1.1870
10	S126	0.0013	1.7520	1.2900
11	S127	0.0471	1.0150	1.5940
12	S128	0.8283	1.1790	0.0000
13	S129	0.0013	0.9278	1.4220
14	S130	0.0171	1.1240	2.0050
15	S131	0.6827	0.9357	1.7630
16	S132	0.0471	0.8698	1.5890
17	PWDM	0.8283	1.1790	0.0000

The discrepancy analysis is depicted in Figure 25, where the high label, mean a high agreement between the expert assessments and the PWDM, the moderate label, means a moderate agreement between the expert assessments and the PWDM and finally, the low label means a low agreement between the expert's assessments and the PWDM.

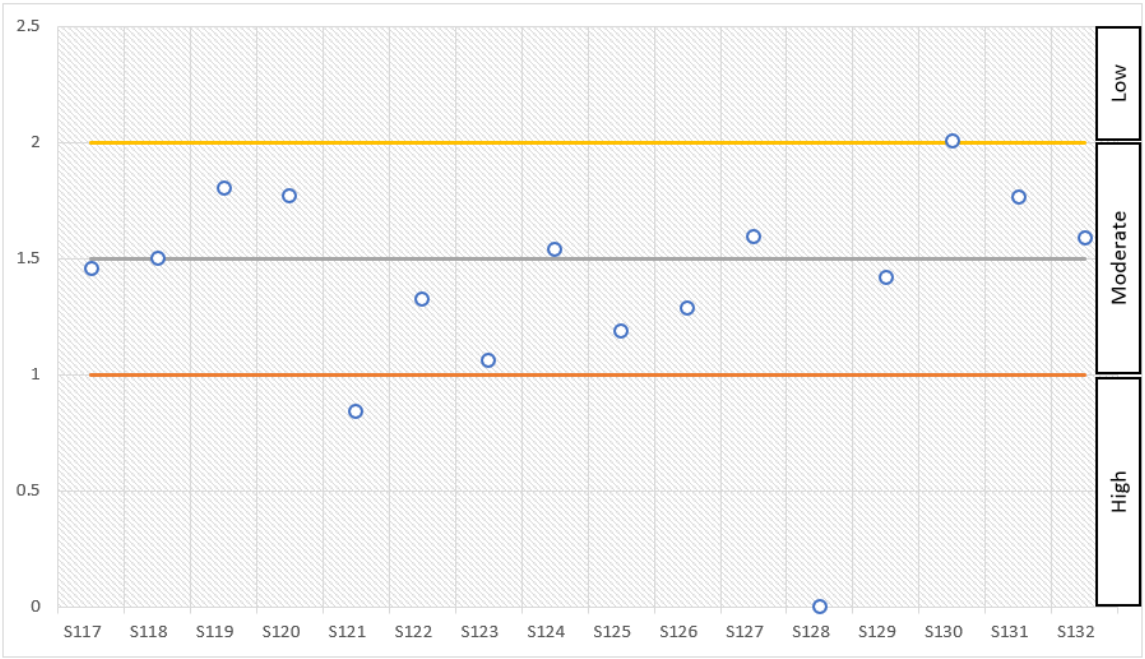


Figure 25 - Discrepancy Analysis Chart

4.3 Bayesian Inference Results

Predicted probability distribution was calculated for all the hardware failure data of widgets A thru D. The Bayes factor (BF) and the Bayesian Information Criterion (BIC) calculations from the summary statistics generated from OpenBUGS will be used as criterions in the hypothesis testing.

Each Table 22, Table 23, and Table 24 is organized in the same manner. The key parameters are as follow:

- Column one shows the code of the widget the data corresponds to.
- Column two shows the node of interest during the BI used in the model script.
- Column three shows the mean value for the node of interest.
- Column four shows the standard deviation of the node of interest predicted values.

- Column five shows the simulation error %.
- Column six shows the number of samples the modeling ran to in order to achieve inference.

The results for the exponential model BI are depicted in Table 22 below. These results account for failure events of widget A thru D as referenced in Table 15.

Table 22 - Exponential Model Summary Statistics for Widgets A thru D

Widget	Node	mean	sd	MC_error	sample
A	BIC	71.98	2.015	0.01859	10000
	BF	0.0432	0.2033	0.00202	10000
B	BIC	86.23	1.722	0.01681	10000
	BF	0.1854	0.3886	0.003686	10000
C	BIC	171	1.343	0.01309	10000
	BF	0.193	0.3947	0.00423	10000
D	BIC	116	1.692	0.01589	10000
	BF	0.0899	0.286	0.00317	10000

The results for the Weibull model BI are depicted in Table 23 below. These results account for failure events of widget A thru D as referenced in Table 15.

Table 23 - Weibull Model Summary Statistics for Widgets A thru D

Widget	Node	mean	sd	MC_error	sample
A	BIC	69.77	2.033	0.02322	10000
	BF	0.2455	0.4304	0.003279	10000
B	BIC	86.35	1.944	0.02807	10000
	BF	0.2903	0.4539	0.003541	10000
C	BIC	171.5	2.018	0.02997	10000
	BF	0.1769	0.3816	0.003155	10000
D	BIC	113.2	2.014	0.02402	10000
	BF	0.3721	0.4834	0.003817	10000

The results for the lognormal model BI are depicted in Table 24 below. These results account for failure events of widget A thru D as referenced in Table 15.

Table 24 - Lognormal Model Summary Statistics for Widgets A thru D

Widget	Node	mean	sd	MC_error	sample
A	BIC	67.49	2.249	0.03562	10000
	BF	0.3008	0.4586	0.004066	10000
B	BIC	83.32	2.331	0.04024	10000
	BF	0.4281	0.4948	0.00496	10000
C	BIC	166	2.073	0.02861	10000
	BF	0.3249	0.4684	0.004211	10000
D	BIC	110	2.125	0.03093	10000
	BF	0.4863	0.4998	0.004996	10000

To facilitate the hypothesis test, the BF results for Widgets A thru D are summarized in Table 25 below.

Table 25 - Bayes Factor Results on Summary Statistics

	Bayes Factor (BF)			Min BF
Widget	Exponential	Weibull	Lognormal	
A	0.0432	0.2455	0.3008	0.0432
B	0.1854	0.2903	0.4281	0.1854
C	0.1930	0.1769	0.3249	0.1769
D	0.0899	0.3721	0.4863	0.0899

Similarly, the BIC results for Widgets A thru D are summarized in Table 26 below. Guidance from (Kelly & Smith, 2011) recommends selecting the minimum value for BIC.

Table 26 - Bayesian Information Criterion (BIC) Results

	BIC			Min BIC
Widget	Exponential	Weibull	Lognormal	
A	72.0	69.8	67.5	67.5
B	86.2	86.4	83.3	83.3
C	171.0	171.5	166.0	166.0
D	116.0	113.2	110.0	110.0

4.4 Hypothesis Testing

The following sections will show hypotheses H1 through H4, both null and alternate, the acceptance criteria and the test conducted on the research data.

4.4.1 Testing H1 and H2 (Program Metrics)

This section is addressing research question RQ1 and hypotheses H1 and H2 shown below for reference.

- RQ1: Is there a cost or time impact when implementing the SEJ process within the DoD risk assessment workflow?
    - **Hypothesis No. 1:**
        - $H1_0$: The average time to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.
        - $H1_a$: The average time to perform a SEJ risk analysis will be less than the traditional risk analysis.
    - **Hypothesis No. 2:**
        - $H2_0$: The average cost to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.
        - $H2_a$: The average cost to perform a SEJ risk analysis will be less than the traditional risk analysis.

Both cost and time measures collected during the research study by the experts during the SEJ are denoted by the Cost-R and Time-R variables, respectively. The cost and time variables are depicted in Table 27 below. The Cost-R variable is the total spent during the SEJ session. The Time-R variable is the sum of the start and end durations for each expert during the SEJ session.

Table 27 - Cost and Time Research Data Collected

Expert	Cost-R ($)	Time-R (hr)
S117	91.00	0.70
S118	116.00	0.97
S119	86.67	0.67
S120	114.00	0.95
S121	66.00	0.60
S122	108.00	0.90
S123	114.00	0.95
S124	60.00	0.50
S125	102.00	0.85
S126	127.83	0.98
S127	138.00	1.15
S128	70.00	0.58
S129	136.50	1.05
S130	112.00	0.93
S131	142.00	1.18
S132	130.00	1.08

Before conducting the Student T-Test both the Cost-R and Time-R datasets were checked for normality to ensure the correct test was done. The resultant normality plots are depicted in Figure 26 for the Cost data and in Figure 27 for the Time data below.

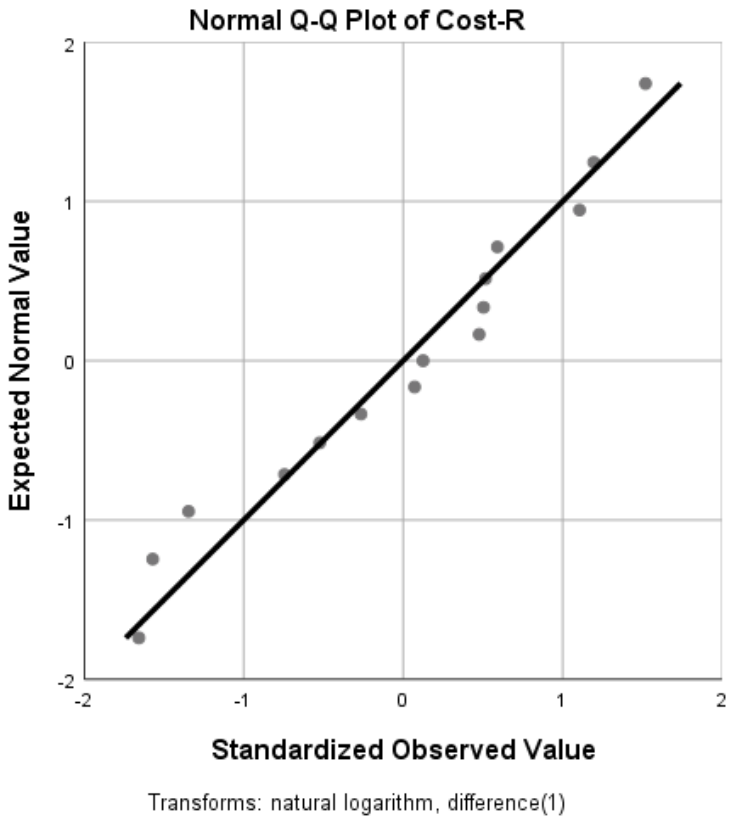


Figure 26 - Normality Plot for Research Cost

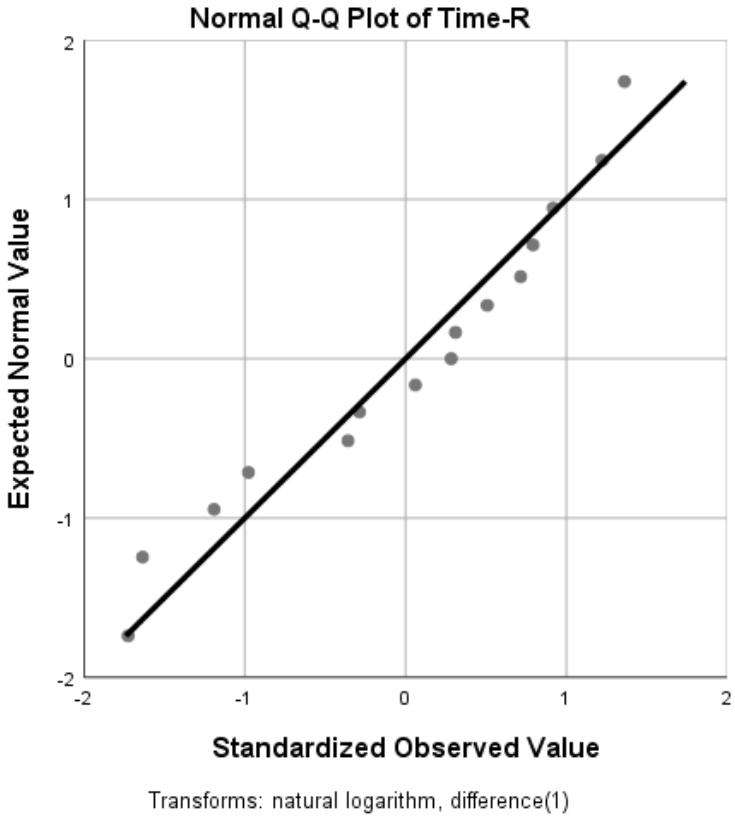


Figure 27 - Normality Test for Research Time

After ensuring both the cost and time data followed a normality test given by the Shapiro-Wilk test statistic and a 95% significance level, then each dataset is tested using a

left tailed Student T-Test one sample statistics. The one sample test statistic uses the mean cost data baselined from projects P1, P2, and P3 for the analysis, as shown in Table 28.

Table 28 – Mean Baseline Cost from P1, P2, and P3

Cost	
Mean Baseline Data P1+P2+P3	$52,080.00

Similarly, the time one sample test statistic uses the mean time data baselined from projects P1, P2, and P3 for the analysis, as shown in Table 29.

Table 29 – Mean Baseline Time from P1, P2, and P3

Time (hours)	
Mean Baseline Data P1+P2+P3	434.00

The results for the one sample statistics for the Time-R dataset are depicted in Table 30 below.

Table 30 - Student T-Test Results for Research Time

One-Sample Statistics						
Research Variable	N	Mean	Std. Deviation	Std. Error Mean		
Time-R	16	0.8781	0.20912	0.05228		
One-Sample Test						
Test Value = 434 hrs.					95% Confidence Interval of the Difference	
Research Variable	t	df	Significance	Mean Difference	Lower	Upper
Time-R	-8284.74	15	2.26E-51	-433.1219	-433.233	-433.01

The results for the one sample statistics for the Cost-R dataset are depicted in Table 31 below.

Table 31 - Student T-Test Results for Research Cost

One-Sample Statistics						
Research Variable	N	Mean	Std. Deviation	Std. Error Mean		
Cost-R	16	107.125	26.048	6.512		
One-Sample Test						
Test Value = $52080					95% Confidence Interval of the Difference	
Research Variable	t	df	Significance	Mean Difference	Lower	Upper
Cost-R	-7981.09	15	3.95E-51	-51972.88	-51986.76	-51959

4.4.2 Testing H3 (SEJ)

This section is addressing research question RQ2 and hypothesis H3, shown below for reference.

- RQ2: Will the statistical significance for the SEJ performance-weighted decision-maker (PWDM) outperform that of a traditional DoD risk assessment?
    - **Hypothesis No. 3:**
        - $H3_0$: The use of the SEJ method will not reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.
        - $H3_a$: The use of the SEJ method will reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.

The statistical accuracy, also known as the *calibration score*, is the p-value of falsely rejecting the hypothesis that the realizations are independently drawn from a distribution complying with the expert's stated percentiles (Cooke, et al., 2014). As defined in Equation 5 a calibration score as defined above equal to 0.05 or higher will suggest that the expert probabilities are statistically supported by the set of seed variables provided (Cooke & Goosens, 2004). Section 4.2 showed the PWDM performance on both the calibration and elicitation phase; these results are summarized in Table 32 below.

Table 32 - Summary of PWDM Performance

	Expert	Calibration Score	Mean Information Score
Expert Perf (Seed Variables)	PWDM	0.8283	1.179
Expert Perf (Target Variables)	PWDM	0.4817	0.4213

4.4.3 Testing H4 (BI)

This section is addressing research question RQ3 and hypothesis H4, shown below for reference.

- RQ3: Will the BI model predictions be enough to outperform that of a traditional DoD risk assessment, based on the available data provided?
    - **Hypothesis No. 4:**
        - $H4_0$: On average there is a difference between the predicted and the observed values when using the Bayes inference probability distribution model.

- $H4_a$: On average there is no difference between the predicted and the observed values when using the Bayes inference probability distribution model.

A summary of the Bayes factor (BF) for all widgets is shown in Table 33 below. The table has marked the resultant acceptance level based on the criteria provided in Table 34. The acceptance criteria to determine the strength of the evidence against the null hypothesis is provided by (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013) in Table 34.

Table 33 - Bayes Factor Acceptance Criteria for Hypothesis Test

Min BF	BF Acceptance Criteria				
	0.01000	0.03125	0.10000	0.31250	1.00000
0.0432			**X**		
0.1854				**X**	
0.1769				**X**	
0.0899			**X**		

Using Equation 22 as the basis of the analysis in OpenBUGS, Table 33 depicts the corresponding minimum Bayes factor as marked by a "*X*", as a result of the guidance provided by (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2013).

Table 34 - BF Acceptance Criteria

BF Acceptance Criteria	
1 to 1/32	Not worth
1/32 to 1/10	Substantial
1/10 to 1/32	Strong
1/32 to 1/100	Very Strong
< 1/100	Decisive

Table 34 shows the different levels of acceptance based on the smallest BF calculated by the predicted model using observed values.

Chapter 5 Conclusions and Recommendations

This chapter will discuss the research findings and conclusions. It will examine the research goals and how these were achieved to reach the results. Also, it will discuss recommendations for practice and future research work.

5.1 Discussion on Research Question RQ1

In section 4.4.1, the results of the program metrics cost and time as a result of implementing the SEJ method within the DoD risk analysis workflow are tabulated and documented. The third research question and the first and second hypotheses are shown below:

- RQ1: Is there a cost or time impact when implementing the SEJ process within the DoD risk assessment workflow?
  - $H1_0$: The average time to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.
  - $H1_a$: The average time to perform a SEJ risk analysis will be less than the traditional risk analysis.
  - $H2_0$: The average cost to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.
  - $H2_a$: The average cost to perform a SEJ risk analysis will be less than the traditional risk analysis.

Normality test is proven by the Shapiro-Wilk test statistic and a 95% significance level, then using a left tailed student t-statistic one sample test to determine the p-value. Using the collected data in Table 27, the results are depicted in Table 30 and Table 31 for time and cost respectively. The resultant p-values for the Time-R variable is 2.25E-51, and

the Cost-R variable is 3.94E-51. In each case the p-value is below the 0.05 acceptance level, suggesting that both null hypotheses can be rejected in favor of the alternate hypothesis. Given the time and cost it took to complete a SEJ session with the current expert pool, as it compares to what it took to perform the same assessment using the traditional method, this suggests there is no cost nor time impact in the SEJ method implementation[1].

5.2 Discussion on Research Question RQ2

As mentioned in section 4.4.2, the statistical accuracy for the SEJ performance-weighted decision-maker (PWDM) is a p-value of 0.05 or higher. Table 32 summarized the results of the PWDM for the seed and target variables and these meet the criteria for statistical accuracy. When examining the expert robustness, the goal is to determine if the loss of any expert could affect the overall results of the SEJ method. Table 19 depicted the results for the expert robustness analysis. The average information score for the original PWDM is calculated at 1.1846, while the average altered PWDM is calculated at 1.1532. As a rule of thumb if the loss of a single expert does not influence the altered PWDM by more than the differences between experts, then the PWDM is responding to genuine variations in expert opinions (Aspinall W. , 2008). The discrepancy analysis results depicted in Figure 25 show that most experts had a high to moderate agreement with the PWDM.

These findings suggest that with the acceptance of both PWDM statistical accuracy and favorable robustness and discrepancy tests the third hypothesis as stated below:

[1] Note: The programmatic findings for RQ1, about cost and time are based on the scope provided by the baseline data from projects P1, P2, and P3. If these parameters are changed in any future SEJ session, the results may change.

- $H3_0$: The use of the SEJ method will not reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.
- $H3_a$: The use of the SEJ method will reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.

The null hypothesis is rejected in favor of the alternate hypothesis based on the findings collected in sections 4.2 and 4.4.2. The research question for H3 stated that:

- RQ2: Will the statistical significance for the SEJ performance-weighted decision-maker (PWDM) outperform that of a traditional DoD risk assessment?

As explained during the elaboration in chapter 2, according to Hubbard and Evans, there are four main biases with scoring methods: cognitive bias, verbal labels, invalid inferences, and invisible correlations (Hubbard & Evans, 2010). When evaluating risks, uncertainty, and probabilities, all practitioners are ineffective. Haselton, Nettle, and Andrews defined this behavior, or cognitive bias, as a systematic pattern deviation from rationality in judgment (Haselton, Nettle, & Andrews, 2005). In other words, there are significant implications to employ scoring methods with cognitive bias to the risk assessment output. These are the primary reasons the traditional risk assessment method is outperformed by the SEJ PWDM, as the PWDM can quantify the uncertainty in the problem given the expert's opinion.

5.3 Discussion on Research Question RQ3

Sections 4.3 and 4.4.3 present the BI results based on the observed data provided by Project P4. There are two metrics with acceptance criteria that will determine the outcome

and optimal model used for probabilistic risk assessments. The second research question and the fourth hypothesis are shown below:

- RQ3: Will the BI model predictions be sufficient to outperform that of a traditional DoD risk assessment, based on the available data provided?
    - $H4_0$: On average there is a difference between the predicted and the observed values when using the Bayes inference probability distribution model.
    - $H4_a$: On average there is no difference between the predicted and the observed values when using the Bayes inference probability distribution model.

The first metric is the BF, which acceptance criteria is depicted in Table 34. The results of the BF measure are depicted in Table 33. This finding suggests a strong to substantial finding in the observed data to reject the null hypothesis in favor of the alternate hypothesis. The second metric, the BIC, will determine the optimal probability distribution. Given the results depicted in Table 26, the lognormal distribution seen to be predicting the probability distribution the closest to the observed values after the model checking. As explained in section 5.1, the current traditional risk assessment method is susceptible to multiple biases and in general, practitioners are ineffective at evaluating the uncertainty in the problem. This is particularly true when observed data is ignored in the risk assessment. A well-informed model using the provided observed data (or evidence) of an event should be able to predict with some uncertainty future events if the model checking metric agrees. Therefore, this finding suggests that the BI method can outperform the current traditional risk assessment.

5.4 Summary of Hypotheses Outcomes

The following section summarizes all the hypotheses and their outcomes in Table 35 below.

Table 35 - Hypothesis Testing Results

Null Hypothesis (H_0)	Alternate Hypothesis (H_a)	Outcome
The average time to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.	The average time to perform a SEJ risk analysis will be less than the traditional risk analysis.	Reject null
The average cost to perform a SEJ risk analysis will be greater than or equal to the traditional risk analysis.	The average cost to perform a SEJ risk analysis will be less than the traditional risk analysis.	Reject null
The use of the SEJ method will not reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.	The use of the SEJ method will reduce the uncertainty from the elicitation of probabilities when compared to the traditional risk analysis.	Reject null
On average there is a difference between the predicted and the observed values when using the Bayes inference probability distribution model.	On average there is no difference between the predicted and the observed values when using the Bayes inference probability distribution model.	Reject null

5.5 Recommendations (Implications) For Practice

As mentioned in the introduction, good risk analysis is the cornerstone of the risk management process. Based on the findings it seems that the implication in practice is that the wrong methodology in the current DoD risk assessment is being used. Given the many flaws already described for the scoring methods currently in place in the DoD risk management process and often the lack technical rigor in the risk analysis exponentially

increases these flaws, it is appropriate to recommend the implementation of both the SEJ and BI method as part of the tools used to asses risk. These methods do not seem to impact critical metrics and can outperform the results of current assessments. The current methodology could be improved as depicted in Figure 2. Practitioners could now use the basic risk model and the methods described in this research to evaluate any subjective or objective risk assessment. Finally, the research study results and methods efficacy should be promoted to key stakeholders within DoD with the intent that one day the policies and instructions that govern the risk management process are changed to incorporate these improvements and best practices within this research.

5.6 Future Work

This research study on introducing a basic risk model to improve the current risk analysis within the DoD risk management process and as a significant finding both the SEJ and BI methods can outperform the current scoring method used for risk analysis. Going forward, it is recommended that practitioners are trained in both the SEJ and BI methods to start a gradual transition from scoring method practices. Ideally, the risk analysis phase can evolve to include processes that can support a community of practice and expert panels at the organization level which could support the level of risk management practitioners and projects need. The first steps of the SEJ method can be done in advance to fulfill the need of assembling an expert panel to review and adjudicate SEJ cases every time. Given the panel could address most topics without the imposed cost burden of the calibration process, given subject matter expertise is not recalibrated once experts are identified and calibrated in their fields.

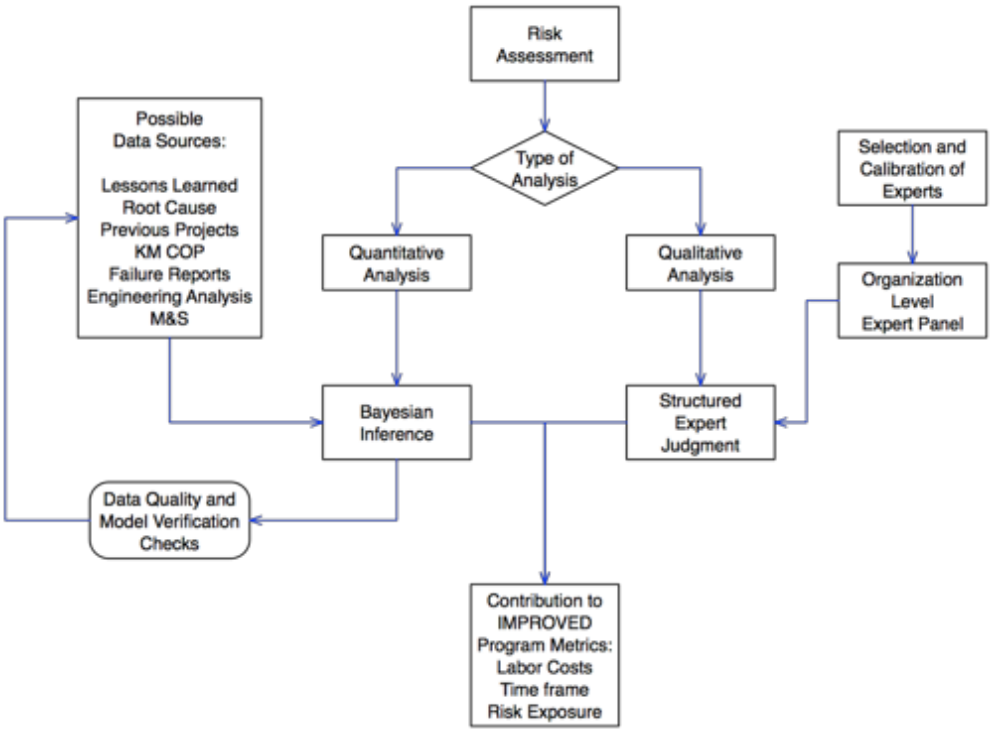


Figure 28- Ideal Framework for Basic Risk Model with Organization Level Support

Future work should explore modeling the Bayes theorem and optimizations of observed data for other probability distributions. The added technical rigor should support a broader range of possible predictive distributions based on observed values using the BIC metric. Also, to further facilitate the collection of observed data for BI analysis, organizations should maintain a central repository with an organization level resources such as lessons learned, failure reports and root cause analysis from previous or current projects. This could aid in the data collection efforts and practitioners could request information from the repository and use the data and evidence to form the basis of their quantitative risk analysis. This ideal framework for the basic risk model is depicted in Figure 28. The following are future recommendations for research based on the findings:

1. Facilitation of an organization level expert panel that could take risk analysis cases from projects and deliver qualitative results without affecting project metrics.
2. Further research can be done to expand further and optimize the BI probability distributions capable of outperforming the current scoring method.
3. Further efforts can be made to help the data collection of observable data that can help the BI analysis.
4. Further research can be done to determine if there is any potential impact on any other program metrics when implementing the SEJ and BI methods. Some program metrics that could be explored are Key Performance Parameter, Key Performance Indicator, and Technical Performance Measures.

References

Aspinall, W. (2008). Expert Judgment Elicitation using the Classical Model and EXCALIBUR. *Briefing notes for Seventh Session of the Statistics and Risk Assessment Section's International Expert Advisory Group on Risk Modeling: Iterative Risk Assessment Processes for Policy Development Under Conditions of Uncertainty / Emerging Infectious Disea*.

Aspinall, W. P. (2006). Structured Elicitation of Expert Judgment for Probabilistic Hazard and Risk Assessment in Volcanic Eruptions. *Stat. Volcanol*, 15-30.

Aven, T., & Heide, B. (2009). Reliability and validity of risk analysis. *Reliability Engineering and System Safety, 94*, 1862–1868.

Ball, D. J., & Watt, J. (2013). Further Thoughts on the Utility of Risk Matrices. *Risk Analysis, 33*(11), 2068–2078.

Baybutt, P. (2017). The Validity of Engineering Judgment and Expert Opinion in Hazard and Risk Analysis: The Influence of Cognitive Biases. *American Institute of Chemical Engineers, 37*(2), 205-210.

Clemen, R. T. (2008). Comment on Cooke's classical method. *Reliability Engineering and System Safety, 93*, 760–765.

Colson, A. R., & Cooke, R. M. (2017). Cross validation for the classical model of structured expert judgment. *Reliability Engineering and System Safety*, 109-120.

Cooke, R. M. (1991). *Experts In Uncertainty, Opinion and Subjective Probability in Science.* New York: Oxford University Press.

Cooke, R. M. (2015). The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? *Risk Analysis, 35*(1).

Cooke, R. M., & Aspinall, W. P. (2011). Quantifying scientific uncertainty from expert judgement elicitation. *Cambridge University Press, 9781107006*.

Cooke, R. M., & Goosens, L. H. (2004). Expert judgement elicitation for risk assessments of critical infrastructures. *Journal of Risk Research*, 643-656.

Cooke, R. M., & Goossens, L. H. (1999). Procedures guide for Structured Expert Judgement. *Delft Univ. Technol. Delft*, 1-50.

Cooke, R. M., & Goossens, L. H. (2008). TU Delft expert judgment database. *Reliability, Engineering, Systems and Safety, 93*, 657–674.

Cooke, R. M., & Jager, E. (1998). A Probabilistic Model for the Failure Frequency of Underground Gas Pipelines. *Risk Analysis*, 511-527.

Cooke, R. M., Wittman, M. E., Lodge, D. M., Rothlisberger, J. D., Rutherford, E. S., Zhang, H., & Mason, D. M. (2014). Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie. *Integrated Environmental Assessment and Management*, 522-528.

Cox, A. L. (2008). What's wrong with risk matrices? *Risk Analysis, 28*(2), 497–512.

DASDSE. (2017). Risk, Issue, and Opportunity Management Guide for Defense Acquisition Programs.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The Development of Progress Plans Using a Performance-Based Expert Judgment Model to Assess Technical Performance and Risk. *Systems Engineering*, 375-391.

Eggstaff, J. W., Mazzuchi, T. A., & Sarkani, S. (2014). The effect of the number of seed variables on the performance of Cooke's classical model. *Reliab. Eng. Syst. Saf, 121*, 72-82.

Haselton, M. G., Nettle, D., & Andrews, P. W. (2005). The Evolution of Cognitive Bias. *The Handbook of Evolutionary Psychology*, 724–746.

Hubbard, D. W. (2009). *The Failure of Risk Management - Why is it broken and how to fix it.* John Wiley & Sons, Inc.

Hubbard, D., & Evans, D. (2010). Problems with Scoring Methods and Ordinal Scales in Risk Assessments. *IBM J. RES. & DEV., 54*(3).

Kahneman, D., & Tversky, A. (1983). Award for Distinguished Scientific Contributions. *American Psychologist*, 2-9.

Kaplan, S. (1997). The Words of Risk Analysis. *Risk Analysis*.

Kaplan, S., & Garrick, J. B. (1991). On The Quantitative Definition of Risk. *Risk Analysis*.

Kelly, D., & Smith, C. (2011). *Bayesian inference for probabilistic risk assessment: A practitioner's guidebook.* Springer.

Kelly, D., Smith, C., Dezfuli, H., Vedros, K., & Galyean, W. (2009). Bayesian Inference for NASA Probabilistic Risk and Reliability Analysis. 204.

Lund, J. R. (2008). A Risk Analysis of Risk Analysis. *J. Contemp. Water Res. Educ., 140*(1), 53-60.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS Book.* New York: Chapman & Hall/CRC Press.

O'Hagan, A., & et.al. (2006). *Uncertain Judgements: Eliciting experts' probabilities.* Wiley.

Pate-Cornell, E. (2002). Finding and fixing systems weaknesses: Probabilistic methods and applications of engineering risk analysis. *Risk Analysis, 22*(2), 319–334.

Peace, C. (2017). The risk matrix: uncertain results? *POLICY AND PRACTICE IN HEALTH AND SAFETY, 17*(2), 131-144.

Project Management Institute. (2017). *PMI Lexicon of Project Management Terms.*

Project Management Institute, Inc. (2017). *A Guide to the Project Management Body of Knowledge* (Vol. Sixth Edition).

Ross, R. S. (2012). Guide for Conducting Risk Assessments. *NIST*.

Ryan, J. J., & et al. (2012). Quantifying information security risks using expert judgment elicitation. *Computers & Operations Research*, 774–784.

Shah, H., & Nowocin, W. (2015). *Engineering Management Body of Knowledge* (Vol. 4th Edition). American Society of Engineering Management.

Smith, E. D., Siefert, W. T., & Drain, D. (2008). Risk Matrix Input Data Biases. *Wiley InterScience, 12*.

Stamatelatos, M., & Dezfuli, H. (2011). Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners.

Thomas, P., & Bratvold, R. B. (2014). The Risk of Using Risk Matrices. *SPE Economics & Management*, 56-66.

Appendix A: Calibration Survey

An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

IRB # 121747

Student Investigator:

Samuel Velazquez, D.Eng. Candidate 540-653-1341

Principal Investigator:

Dr. Thomas Mazzuchi, Sc.D., EMSE Dept. GWU 202-994-9187

The George Washington University

Engineering Management and Systems Engineering Department

INSTRUCTIONS: *Use* Table A - 1 *as a reference to aid your determination of each calibration question's subjective quantile value. Answer each question* <u>individually</u> *for Sections 1 through 3. The Student Investigator will clarify any questions regarding the meaning of a specific term.*

Table A - 1 – Risk Exposure Tool (Calibration Survey)

Risk Exposure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	Minimal impact to the project					Minor impact to the project					Moderate impact to the project					Significal impact to the project					Critical impact to the project				

Item #1

Start Time:	

<u>Description</u>:

The sponsor tasked the government's Spartan Team to develop a widget with a final delivery date of 5 months for today; however, the widget's requirements are currently

known only to the sponsor. A critical component is that pending production cannot proceed until the Spartan team reviews the manufacturing package delivered by the vendor. The sponsor has set strict delivery milestones for all components and tests in the future.

The team must review over 200 documents in a short, approximately three-week timeframe. The systems engineering team will need to form three five-person sub-teams to review these documents so that each team reviews ten documents per week.

Documents in this case can be Risk Management Plan (RMP), Configuration Management Plan (CMP), Interface Requirements Specification (IRS), Interface Control Document (ICD), System Requirements Specification, System/Subsystem Specification (SSS), Contract Data Requirements List (CDRL) or a combination.

Statement:

Since a 4-month production lead time is expected, if Spartan Team cannot accommodate the review of the documents while also maintaining the component development schedule, then milestones and deliverables will be delayed.

Based on your experience, estimate the likelihood that the Spartan sub-teams to be able to finish on time and be able to release the widget for production as planned. Please provide your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 2 – Calibration Answer Item #1

5%	50%	95%	Rationale

Item #2

Description:

A government Spartan team has been tasked to assess the capabilities of a new widget that just finished production with only a small, fixed budget. All the widget's capabilities are listed in document [XPC-011], but its primary requirement is an underwater test with a special "adapter" to be able to withstand severe shock. Unfortunately, the adapter needs to be built first because it is unique for every task.

The pool vendor contracted by the Spartan team has a small, two-to-three-week window for the capability test. The adapter vendor can expedite the manufacturing, but history shows the adapter quality varies with expedited deliveries as shown in the chart below:

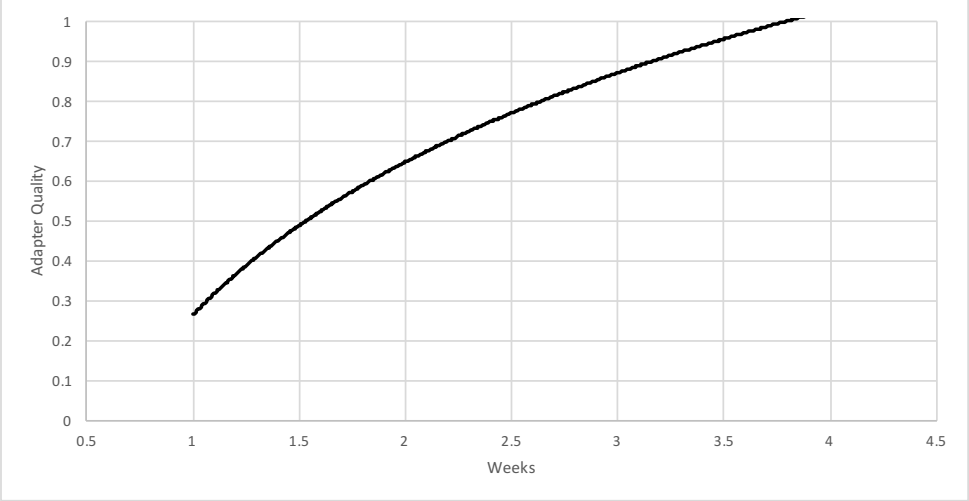


Figure A - 1 – Adapter Quality Over Time

Statement:

If the Spartan team does not complete the widget test in the pool within the given timeframe due to the adapter's quality or availability issues, then the widget will not satisfy the sponsor's requirements. Based on your experience, estimate the possible impact to the test based on the uncertainty in the adapter's quality based on delivery and the effects this

may have on the widget shock test. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 3 – Calibration Answer Item #2

5%	50%	95%	Rationale

Item #3

Description:

A government Spartan team has been tasked to assess the impact of a late funding request to the program office to support the next phase of the widget development after the widget unexpectedly survived its shock trials. The next phase of the widget development requires the development of new software architecture. This positions the sponsor in a risky political position with Congress, especially now that "funds are strictly controlled.". There is some degree of uncertainty for funds being awarded. Also, the Spartan team has a hard deadline to make the assessment before the end of the Fiscal Year (FY) 2017 (assume we are still in Aug 2017), so approximately 3-4 weeks.

Statement:

If the Spartan team does not complete the assessment, and the widget's new software architecture funds are not awarded before the end of the FY, then the program office will not be recommending the award for the next FY.

Based on your experience, estimate the funding award uncertainty for the widget based on the given scenario and timeline limitations. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 4 – Calibration Answer Item #3

5%	50%	95%	Rationale

Item #4

Description:

A government Spartan team has been tasked to troubleshoot a capability implementation to the new widget software. The new pixilation capability is described by the charts below, and its usage depends on the level set. Each level uses additional processing power and time based on its CPU. The sponsor only to procure Panntium 74.9 CPUs (not to be confused with that other chip manufacturer).

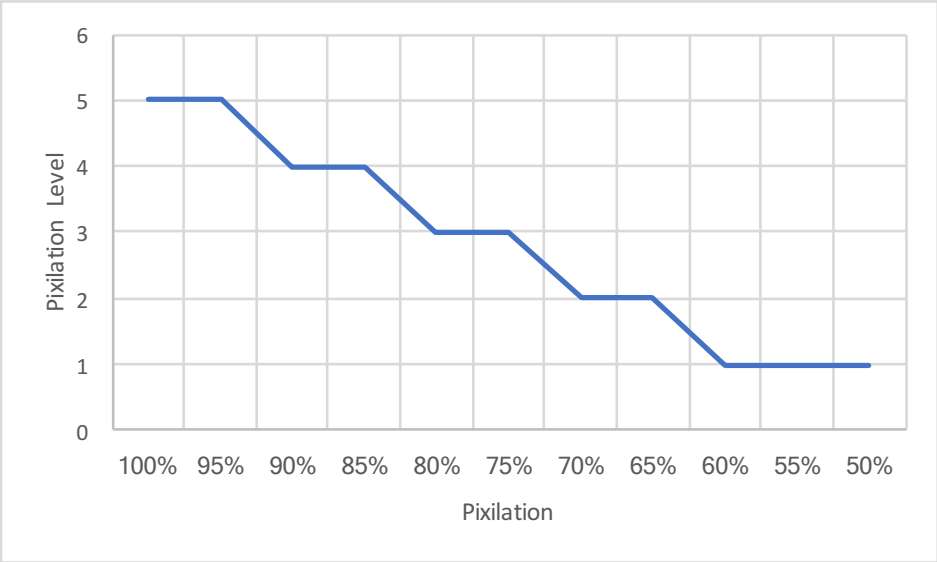


Figure A - 2 – Pixilation Levels

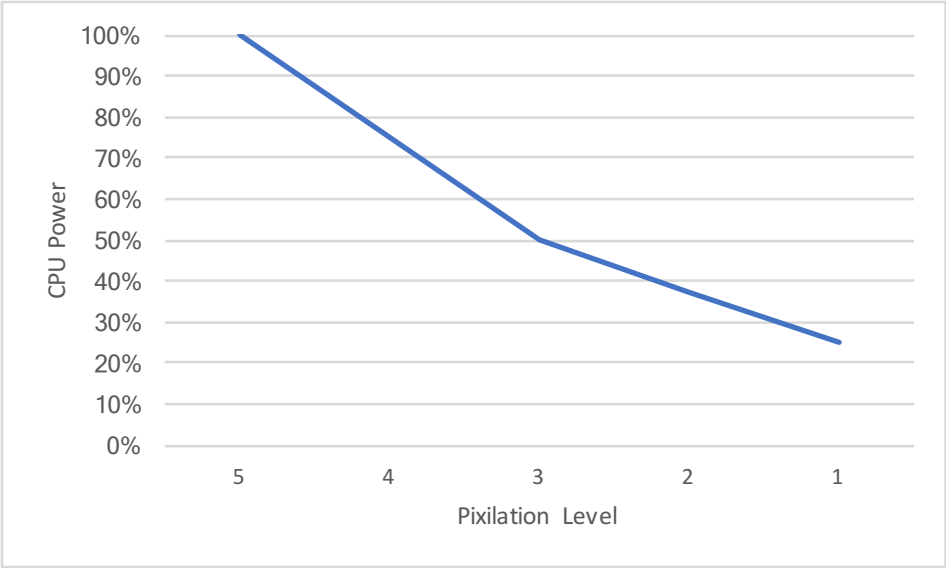


Figure A - 3 – CPU Power

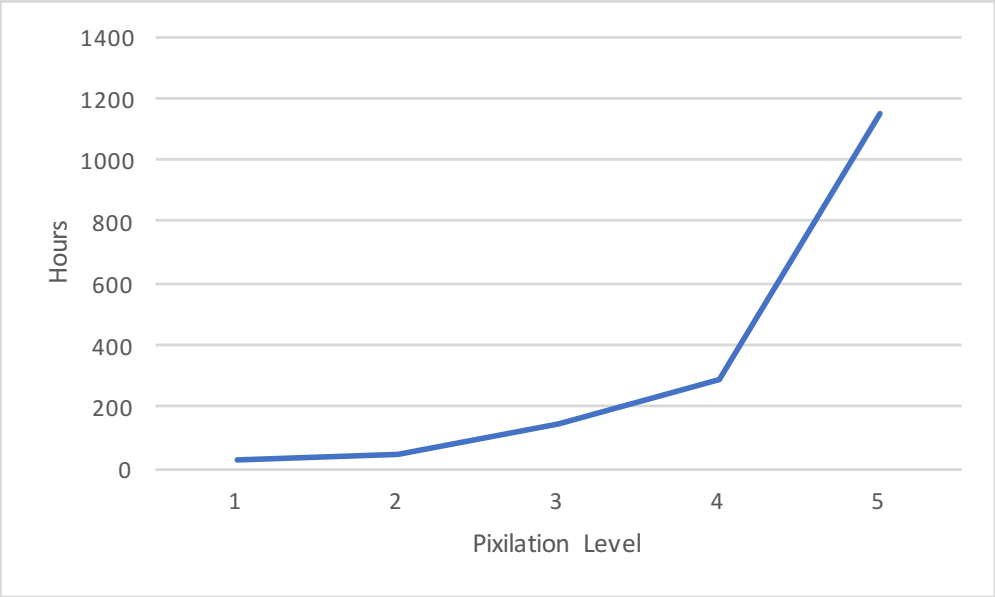


Figure A - 4 – Pixilation Level Over Time

The sponsor has requested this capability be added to the widget controlling software since it had to justify to Congress why funding was needed for that other software last FY. In return, he recommended to the Spartan team lead that the pixilation computational time

should take between 1-2 weeks and should try to achieve the highest pixilation possible. It is important to note the Panntium CPUs have an 80% reliability rate.

Statement:

If this widget pixilation capability is not implemented according to the sponsor's requirements, then the sponsor cannot support its customers by providing the needed product in time. Based on your experience and the scenario above, estimate the uncertainty for the widget' capability implementation for the pixilation software in the time and level requested by the sponsor, considering the CPUs reliability. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 5 – Calibration Answer Item #4

5%	50%	95%	Rationale

Item #5

Description:

A government Spartan team has been tasked to assess the project's compliance with the Risk Management Framework (RMF) milestones. These milestones need to be addressed early in the program to allow time for proper implantation of Information Assurance and Anti-tamper mitigation plans. Due to the program's aggressive schedule, and the length of time it generally takes to produce the RMF documentation, there is a chance that not all necessary documentation and assessments will be completed. For the estimates, assume the following:

- Each document's typical approval timeline is two-months,

- project schedule has no slack,
- The project has some scope creep, and
- The project has 50% leveled on all resources assigned.

Statement:

If the program RMF requirements among these (Information assurance and Anti-tamper) are not met, then the program milestones and mitigation plans will not be addressed, and RMF certification will not be given to the program to continue to their next phase.

Based on your experience, and using the scenario above, estimate the uncertainty to draft the two RMF requirements documents within the given program schedule assumptions. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 6 – Calibration Answer Item #5

5%	50%	95%	Rationale

Item #6

Description:

A local vendor is about to integrate two "components" that have been under development for several years. Component one's schedule ends in about two months and Components 2 ends in about 1.5 months. There is high uncertainty and anxiety regarding the integration since the sponsor wants to test the newly integrated Component-X in 4 weeks. The entire development timeline has taken about three years, and the sponsor needs a functioning, integrated Component-X. A government Spartan team will focus on trying

to reduce any potential risks and reduce the schedule conflict by parallel schedule tasking. The team will monitor any potential risk-adders due to the schedule crunch efforts that may increase the integration risks on either component beyond acceptable levels.

Assumptions:

1. Will move manpower from Component 2 to Component 1 to decrease timeline,
2. Will increase working hours by providing O/T and voluntary Saturday work, and
3. Will monitor production line for waste.

Statement:

If the Component-X integration is unsuccessful due to any potential schedule and risk reduction efforts to meet the sponsor's timeline, then the sponsor will consider this effort a failure since Component-X was needed to support platform integration of Component-YX.

Based on your experience, and using the scenario above, estimate the uncertainty of delivering Component-X on-time to the sponsor. Use the given assumptions, but do not limit the scenario to these as you can generate others based on your experiences. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 7 – Calibration Answer Item #6

5%	50%	95%	Rationale

Item #7

Description:

A government Spartan team has been tasked to review the requirements of a program with three shared components with similar traits. Unfortunately for the project lead, the

requirements are changed on a monthly basis by the component owners, and there is little-to-no CM in the program. The sponsor has made it clear that, in order to succeed, the team must achieve requirements alignment.

Assumptions:

1. The requirements have changed for the last six months,
2. The establishment of a reliable CM program will take at least four months,
3. A review of each component requirement and approval takes about one month (there is some commonality), and
4. The sponsor has requested a delivery of the components eight months from today.

Statement:

If the requirements for all three components are not aligned, then the delivery schedule set by the sponsor will not be met (i.e., further funding will stop). Based on your experience, and using the scenario above, estimate the uncertainty to prevent the funding shortage to the program due to requirements misalignment. Please give your 5%, 50% and 95% quantile estimate and rationale:

Table A - 8 – Calibration Answer Item #7

5%	50%	95%	Rationale

Item #8

Description:

A government Spartan team has been tasked to assess a possible delay to obtain the materials required to build the system. The manufacturing and purchasing plan shows $5M in purchasing costs for raw materials. Unfortunately, due to delays in contracting the sub-

vendor that provides the Ubnatium-123 ore, purchases orders are in limbo. Even worse, the program contracting officer decided to quit the company to pursue his dream job.

Assumptions:

1. Some preliminary design review has occurred,
2. The design is ready for the critical design review, and
3. There is high urgency from the sponsor.

Statement:

If the order for raw materials experiences further delays, then delivery it can jeopardize the delivery of the "system" design and prevent from achieving the set milestones

Based on your experience, and using the scenario above, estimate the uncertainty of placing the raw material orders in time to prevent or miss any key milestone in the program. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 9 – Calibration Answer Item #8

5%	50%	95%	Rationale

Item #9

Description:

A government Spartan team has been tasked to investigate the rate at which test documentation is being completed for the project in question. An outside oversight entity has recommended that 50% of the documentation be completed ahead of the scheduled timeline to meet program demands.

Assumptions:

For this item, assume the following:

1. The engineering team works 8hrs/day and 40hrs/week,
2. The number of test documents is between 100-125,
3. The rate of completion is 1 document per week, and
4. The Sponsor completion timeline is not specified.

Statement:

If the completion and delivery of all test documents are delayed, then the test program and the integration of the hardware package will be deferred, and further install will be missed.

Based on your experience, and using the scenario above, estimate the delivery of the test documents with the uncertainty of program requirements. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 10 – Calibration Answer Item #9

5%	50%	95%	Rationale

Item #10

Description:

A government Spartan team has been asked to calibrate a model with historical data, which will be used for verification of the widget. Unfortunately, the historical data has gaps, and the team has not decided on the calibration method yet, due to its complexity. The sponsor suggested that calibration data can be obtained during the scheduled verification test in 2 months. Ultimately, the team should be able to support a plan that meets the "Widget-Z" milestones. During verification testing, the team was not able to obtain all data. Just 25% of the planned data was captured due to the failure of the test article. The sponsor communicates his concerns over the historical data gaps and newly

collected data shortcomings and the uncertainty this could cause to the calibration of the model.

Statement:

If the team cannot properly calibrate the model w/ the data at hand, then widget-z requirements will not be met. In your experience and using the scenario above estimate the uncertainty of the model calibration based on the estimated data available and impact to widget-z milestone/requirements. Please give your 5%, 50% and 95% quantile estimate and rationale below:

Table A - 11 – Calibration Answer Item #10

5%	50%	95%	Rationale

End Time:	

Appendix B: Elicitation Survey

An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

IRB # 121747

Student Investigator:

Samuel Velazquez, D.Eng. Candidate 540-653-1341

Principal Investigator:

Dr. Thomas Mazzuchi, Ph.D., EMSE Dept. GWU 202-994-9187

The George Washington University

Engineering Management and Systems Engineering Department

INSTRUCTIONS: Use the Table B - 1 as a reference to aid your determination of the uncertainty value for each question in the quiz. Record the start and end times. Most assumptions have been provided; some need to be inferred. Clarification questions should be directed to the Student Investigator.

Table B - 1 – Risk Exposure Tool (Elicitation Survey)

Risk Exposure	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
	Minimal impact to the project					Minor impact to the project					Moderate impact to the project					Significal impact to the project					Critical impact to the project				

Individual Elicitations

Item #1

Start Time:	

The **target variable** is defined below:

Widget malfunction: Typically, the widget functions its optimal mechanical functional design parameters by moving from an open and closed position. Assume the widget material is aluminum. Its normal thermal loading capacity is 1000 cycles at room temperature.

The **query variable** is defined below:

Thermal load failure (simplified): The failure type the widget will experience in an *uncertain event* due to higher-than-normal usage exceeding room temperature.

Description

A government Spartan team has been tasked to evaluate a new widget prototype. The widget has a higher-than-normal thermal loading on the widget can be associated to the increase of usage at a high rate or a high number of locks between usages, which could lead to a loss of structural integrity. The mission requirements dictate the widget functions for a total of 10,000 at 200 F.

Assumption:

Wear and tear on the widget are a linear function between the temperature and the cycles as shown in the chart. Room temperature is defined as 50F +/- 5F.

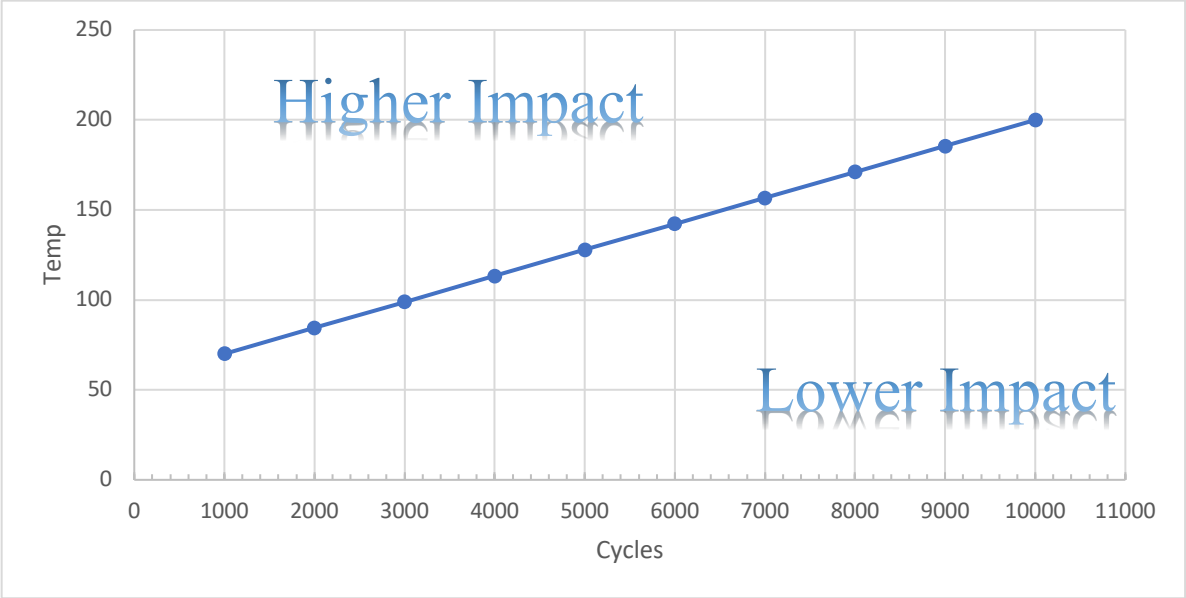


Figure B - 1 – Widget Temperature

Statement:

If the widget malfunctions due to thermal load failure caused during its operating cycles (higher than average usage) at 5,000 cycles and 150 F, then this can lead to the prototype not being accepted and eventual project failure.

Based on your experience, and using Table B - 1 as guidance, estimate the uncertainty of a widget malfunction due to thermal failure gave the higher-than-normal usage conditions as a RE to the Spartan team's evaluation of the widget prototype. Please give your answers in 5%, 50% and 95% quantile values and provide a rationale below:

Table B - 2 – Elicitation Answer Item #1

5%	50%	95%	Rationale

Item #2

The **target variable** is defined below:

Reliability: Theoretically defined as ***1 - "probability of failure;"*** or, in terms of availability, {uptime/ (uptime + downtime)}

The **query variable** is defined below:

Component Long-Term Storage: The length of time (*uncertain event*) that a commodity may be stored without becoming unfit for use.

Description:

A government Spartan team has been tasked to evaluate the reliability impact of new Hybrid Hard Drive (HDD) prototype based on and older HDD model focusing on long-term storage. In certain documented cases, there can be mechanical or electrical degradation as a result of long-term storage on HDDs. The team was provided historical data of the old model as shown in the chart below. A system using the old HDDs has been operating for about 30 years non-stop and will see a new HDDs replacement with the new prototypes. These prototypes have been in storage for about 29 years.

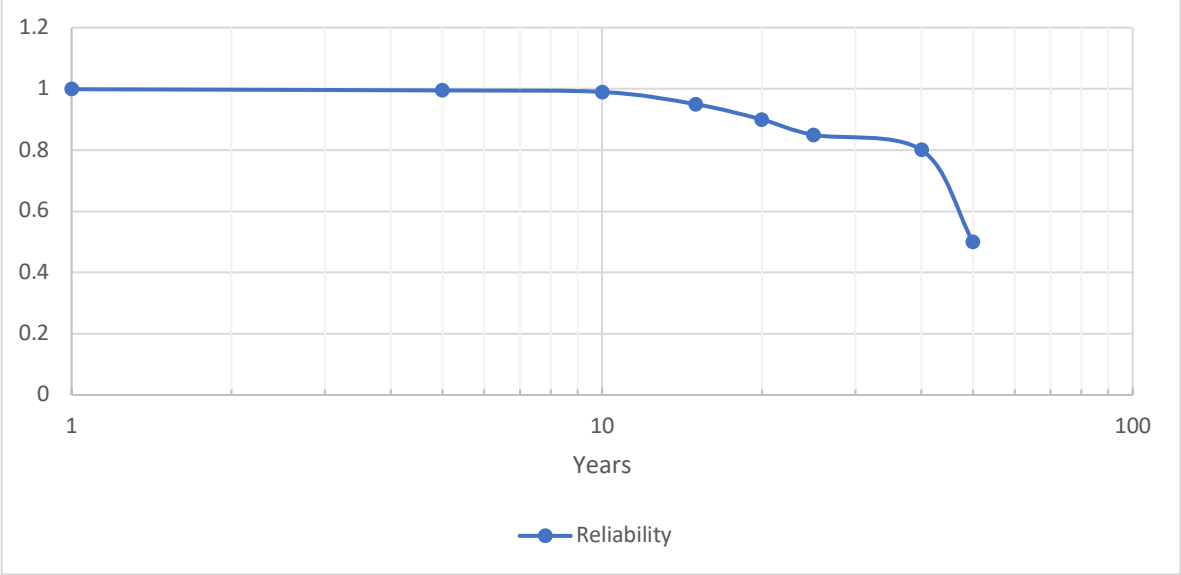


Figure B - 2 – Electrical Degradation of Older HDD Model

Assumptions:

It is estimated that the new prototype has the same or better reliability of the old model up to 10 years in long storage. The engineers estimate a 25% improvement between 10 and 25 years; however, there is a cone of uncertainty of 50 years at about 50% reliability. The new HDD reliability seems to be trending linearly.

Statement:

If the system is impacted due to the new HDD's reliability, then all cloud storage capabilities will be lost. Based on your experience, and using Table B - 1 as guidance, estimate the uncertainty of the new HDD's reliability based on its long-term storage. Please give your answers in 5%, 50% and 95% quantile values and provide a rationale below:

Table B - 3 – Elicitation Answer Item #2

5%	50%	95%	Rationale

Item #3

The **target variable** is defined below:

Communication: Leadership misinformation may impact communication and increase the level of uncertainty.

The **query variable** is defined below:

Delivery Time Frame: In this instance, the delivery timeline causes the *uncertain event*.

Description:

A government Spartan team has been tasked to determine the cause of a possible delay of an important government program. There is a high level of anxiety due to recent delays by the current vendor working a contract modification. The vendor is currently assembling the last components for a test article of a system of widgets that will undergo a

demonstration in front of the user. The team is collecting data from all key team members to determine the status of the assembly.

The chief engineer has quoted two weeks to completion. The engineers working the project have quoted 2.5 weeks to completion. The project lead estimates one week to complete. The president of the vendor company makes states that they are on target to delivery.

Assumption:

The government milestone for delivery of the test article is at least 1-2 weeks more than what the vendor's engineers quoted.

Statement:

If the delivery of the test article modification is delayed then, the final system delivery demo will be delayed, and possibly canceled, which will damage relations with the end user.

Based on your experience, and using Table B - 1 as guidance, estimate the level of uncertainty regarding risk exposure due to all levels of miscommunication (in this case) that could have impacted the delivery of the test article. Please give your answers in 5%, 50% and 95% quantile values and provide a rationale below:

Table B - 4 – Elicitation Answer Item #3

5%	50%	95%	Rationale

Item #4

The **target variable** is defined below:

Technical Performance Measure (TPM): A prediction and demonstration of the degree of an anticipated or actual achievement of selected technical objectives.

The **query variable** is defined below:

Incremental Funding: In this instance, incremental funding introduces an *uncertain event* to the situation due to the needs of the program office.

Description:

A government Spartan team has been asked to support a planned test for an important project. The program office is currently holding incremental funding due to high uncertainty on the latest performance reports coming from the development team. The last integrated master schedule (IMS) review puts the test in the critical path, and further delays will negatively impact the project execution.

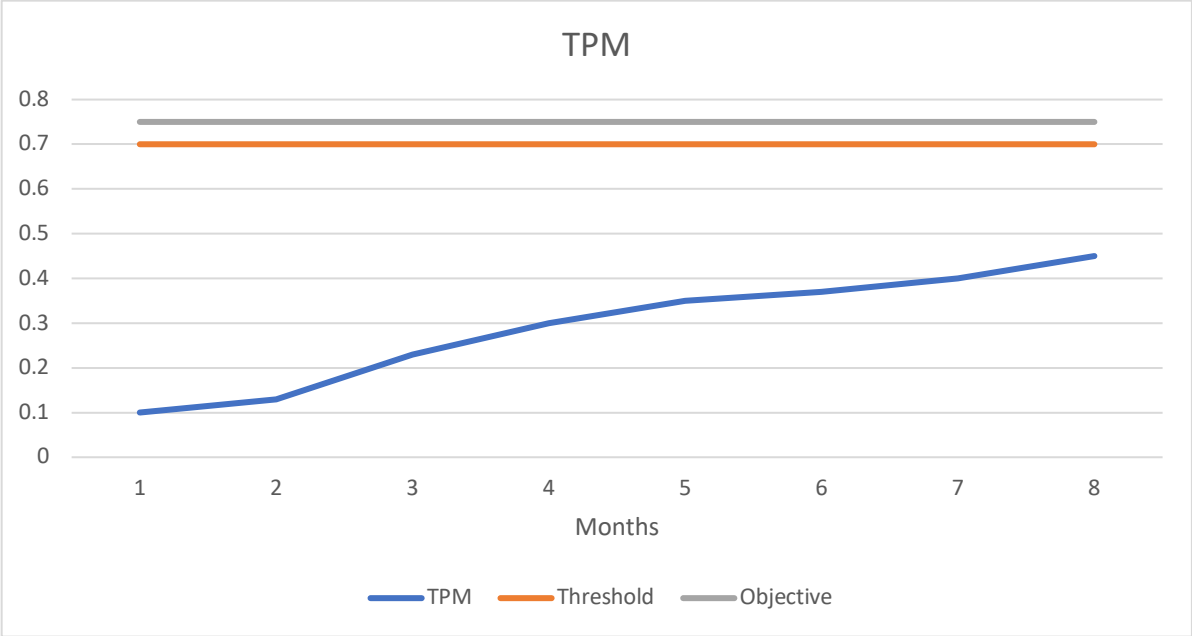


Figure B - 3 – TPM Chart

Assumption:

1. The test is scheduled for month #7, and
2. The development team can achieve threshold, at some point shortly.

Statement:

If the development team does not achieve the desired TPM performance in a timeframe consistence with the IMS critical path, then incremental funding will not be released to support project testing. Based on your experience, and using Table B - 1 as guidance, estimate the uncertainty based on the demonstrated performance from the development team as shown on the TPM chart and the assumption the incremental funds will be released by the program office without impacting the test. Please give your answers in 5%, 50% and 95% quantile values and provide a rationale below:

Table B - 5 – Elicitation Answer Item #4

5%	50%	95%	Rationale

Appendix C: Consent Form

An Improved Risk Analysis Methodology Focused on Expert Elicitation and Probabilistic Assessments to the Risk Management Process of the Department of Defense

IRB #121747

Student Investigator: Samuel Velazquez, D.Eng. Candidate 540-653-1341

Principal Investigator: Dr. Thomas Mazzuchi, Ph.D., EMSE Dept. GWU 202-994-9187

You are invited to take part in a research study being conducted

by Samuel Velazquez D.Eng. Candidate at

The George Washington University,

Engineering Management and Systems Engineering Department.

***Section 1** – The basic qualifications for the research study include the following:*

- *Ability to effectively communicate,*
- *Familiar with basic concepts of risk analysis, and*
- *Familiar with the representation of uncertainty as a subjective probability.*

Please read this form and ask me any questions that will help you decide if you want to be in the study. Taking part is entirely voluntary, and, even if you decide you want to participate, you can quit at any time. Your employment status will not be affected in any way should you choose not to take part or choose to withdraw at any time.

***Section 2** - The study will try to answer the following research questions:*

1. *Will a structured expert elicitation method be able to replace the current assessment methods that lead to the 5x5 matrix?*
2. *Could probabilistic risk assessment be incorporated into the risk analysis phase and yield better cost benefits during a risk implementation?*

3. *Is the cost of a structured expert elicitation method justified?*

Section 3 *- The total amount of time you will spend in this study is 60 minutes each week for 12 weeks. All research participants will assist in the risk assessment for select topics using a structured elicitation method. The following steps will be followed in this method:*

1. *Preparation for elicitation:*
    - *Definition of the case structure*
    - *Identification of target variables*
    - *Identification of query variables*
    - *Identification of performance variables*
    - *Identification of experts*
    - *Selection of experts*
    - *Definition of elicitation format document*
    - *Dry run exercise*
    - *Expert training session*
2. *Elicitation*
    - *Expert elicitation sessions*
3. *Post-elicitation*
    - *Combination of expert assessments*
    - *Discrepancy and robustness analysis*
    - *Feedback*
    - *Post-processing analyses*
    - *Documentation*

Section 4 *- The study has the following risks:*

- *This is a non-funded, voluntary effort; hence, this implies that the research participant will carry some financial burden (or risk) while attending this study since it will be performed during working hours.*

Section 5 *- The study will request from the research study volunteers the following information for the following purpose:*

- *Name – The name will be used to assign a unique expert identifier for the research study, e.g., Name1 = E1*
- *Grade – Will be used to gather information towards the hypothesis. The grade measure will give an estimate for how much labor costs/Experts is spent.*
- *Field of expertise – Will be used to better match the elicitation questions. No more than 3 per expert.*

Taking part in this research will train you on the benefits of expert elicitation, improve your understanding of how to evaluate project benefits using effective risk metrics, and provide experience in project technical risk evaluation in a collaborative environment.

You can contact the Student Investigator listed on the front of this form at 540-653-1341, if you have questions, concerns, complaints, or think you have been harmed. For questions regarding your rights as a participant in human research call the GWU Office of Human Research at 202-994-2715.

If you agree to take part in this study, please sign below:

Research Participant Signature *Date*

After you sign this consent form, the Student Investigator will provide you with a copy. Please keep this copy in case you want to reread it or call someone about the study.

Appendix D: Consent Form

Table D - 1

		QUANTILES		
EXPERTS	**VARIABLE**	**0.05**	**0.50**	**0.95**
S117	FINISH_ON_TIME	6	14	18
S117	ADAPTER_QUALIT	9	15	21
S117	FUNDING_AWARD	18	22	24
S117	CAPABILITY_IMP	3	6	16
S117	DRAFT_REQUIREM	11	15	23
S117	DELIVERY_COMPO	16	22	25
S117	PREVENT_FUNDIN	12	18	22
S117	PLACING_RAW_MA	10	18	23
S117	PROGRAM_REQUIR	2	4	8
S117	MODEL_CALIBRAT	6	10	17
S117	WIDGET_MALFUNC	16	21	24
S117	RELIABILITY	12	13	15
S117	COMMUNICATION	6	9	13
S117	TECHNICAL_PERF	19	22	25
S118	FINISH_ON_TIME	6	18	21
S118	ADAPTER_QUALIT	15	18	25
S118	FUNDING_AWARD	15	17	25
S118	CAPABILITY_IMP	8	10	12
S118	DRAFT_REQUIREM	4	16	22
S118	DELIVERY_COMPO	16	21	25
S118	PREVENT_FUNDIN	5	12	20
S118	PLACING_RAW_MA	15	20	23
S118	PROGRAM_REQUIR	3	5	7
S118	MODEL_CALIBRAT	11	18	24
S118	WIDGET_MALFUNC	8	12	20
S118	RELIABILITY	8	12	22
S118	COMMUNICATION	5	10	16
S118	TECHNICAL_PERF	13	19	23
S119	FINISH_ON_TIME	15	19	23
S119	ADAPTER_QUALIT	11	13	16

S119	FUNDING_AWARD	19	22	24
S119	CAPABILITY_IMP	10	15	21
S119	DRAFT_REQUIREM	5	9	13
S119	DELIVERY_COMPO	20	22	25
S119	PREVENT_FUNDIN	6	10	16
S119	PLACING_RAW_MA	5	9	14
S119	PROGRAM_REQUIR	7	11	16
S119	MODEL_CALIBRAT	4	7	11
S119	WIDGET_MALFUNC	17	21	25
S119	RELIABILITY	8	13	17
S119	COMMUNICATION	10	14	17
S119	TECHNICAL_PERF	18	21	24
S120	FINISH_ON_TIME	18	21	24
S120	ADAPTER_QUALIT	13	16	18
S120	FUNDING_AWARD	12	13	14
S120	CAPABILITY_IMP	20	21	22
S120	DRAFT_REQUIREM	8	11	13
S120	DELIVERY_COMPO	10	16	18
S120	PREVENT_FUNDIN	18	19	20
S120	PLACING_RAW_MA	6	8	10
S120	PROGRAM_REQUIR	13	14	15
S120	MODEL_CALIBRAT	4	6	8
S120	WIDGET_MALFUNC	23	24	25
S120	RELIABILITY	2	4	6
S120	COMMUNICATION	6	11	20
S120	TECHNICAL_PERF	14	16	18
S121	FINISH_ON_TIME	5	12	23
S121	ADAPTER_QUALIT	7	15	25
S121	FUNDING_AWARD	9	17	23
S121	CAPABILITY_IMP	5	15	23
S121	DRAFT_REQUIREM	6	14	22
S121	DELIVERY_COMPO	8	16	22
S121	PREVENT_FUNDIN	6	16	23
S121	PLACING_RAW_MA	6	15	25
S121	PROGRAM_REQUIR	15	18	22
S121	MODEL_CALIBRAT	10	16	25
S121	WIDGET_MALFUNC	14	18	23

S121	RELIABILITY	12	15	18
S121	COMMUNICATION	10	15	20
S121	TECHNICAL_PERF	11	16	23
S122	FINISH_ON_TIME	12	17	21
S122	ADAPTER_QUALIT	15	18	21
S122	FUNDING_AWARD	8	13	18
S122	CAPABILITY_IMP	1	6	11
S122	DRAFT_REQUIREM	7	10	13
S122	DELIVERY_COMPO	21	23	25
S122	PREVENT_FUNDIN	9	13	17
S122	PLACING_RAW_MA	18	21	24
S122	PROGRAM_REQUIR	1	8	15
S122	MODEL_CALIBRAT	11	16	21
S122	WIDGET_MALFUNC	11	16	21
S122	RELIABILITY	16	18	20
S122	COMMUNICATION	6	9	23
S122	TECHNICAL_PERF	16	18	20
S123	FINISH_ON_TIME	12	18	23
S123	ADAPTER_QUALIT	12	18	24
S123	FUNDING_AWARD	4	11	21
S123	CAPABILITY_IMP	2	9	13
S123	DRAFT_REQUIREM	6	13	18
S123	DELIVERY_COMPO	18	20	25
S123	PREVENT_FUNDIN	6	11	15
S123	PLACING_RAW_MA	16	21	24
S123	PROGRAM_REQUIR	1	13	25
S123	MODEL_CALIBRAT	10	14	18
S123	WIDGET_MALFUNC	15	19	24
S123	RELIABILITY	11	18	23
S123	COMMUNICATION	2	7	15
S123	TECHNICAL_PERF	11	18	23
S124	FINISH_ON_TIME	13	19	22
S124	ADAPTER_QUALIT	11	19	21
S124	FUNDING_AWARD	2	20	24
S124	CAPABILITY_IMP	7	9	13
S124	DRAFT_REQUIREM	12	14	18
S124	DELIVERY_COMPO	13	19	22

S124	PREVENT_FUNDIN	19	22	24
S124	PLACING_RAW_MA	13	19	22
S124	PROGRAM_REQUIR	7	9	13
S124	MODEL_CALIBRAT	13	19	22
S124	WIDGET_MALFUNC	11	13	15
S124	RELIABILITY	18	20	21
S124	COMMUNICATION	19	21	22
S124	TECHNICAL_PERF	12	14	16
S125	FINISH_ON_TIME	5	20	25
S125	ADAPTER_QUALIT	6	15	20
S125	FUNDING_AWARD	9	18	23
S125	CAPABILITY_IMP	8	13	18
S125	DRAFT_REQUIREM	7	17	22
S125	DELIVERY_COMPO	5	10	15
S125	PREVENT_FUNDIN	10	18	22
S125	PLACING_RAW_MA	9	17	21
S125	PROGRAM_REQUIR	5	15	19
S125	MODEL_CALIBRAT	9	18	24
S125	WIDGET_MALFUNC	10	15	25
S125	RELIABILITY	9	19	23
S125	COMMUNICATION	15	20	24
S125	TECHNICAL_PERF	5	10	15
S126	FINISH_ON_TIME	17	20	22
S126	ADAPTER_QUALIT	19	21	24
S126	FUNDING_AWARD	10	15	20
S126	CAPABILITY_IMP	17	20	22
S126	DRAFT_REQUIREM	20	22	25
S126	DELIVERY_COMPO	10	14	18
S126	PREVENT_FUNDIN	16	20	21
S126	PLACING_RAW_MA	7	11	18
S126	PROGRAM_REQUIR	14	19	21
S126	MODEL_CALIBRAT	14	16	22
S126	WIDGET_MALFUNC	5	10	15
S126	RELIABILITY	5	8	10
S126	COMMUNICATION	10	15	20
S126	TECHNICAL_PERF	10	15	20
S127	FINISH_ON_TIME	5	10	15

S127	ADAPTER_QUALIT	15	20	25
S127	FUNDING_AWARD	10	15	20
S127	CAPABILITY_IMP	10	15	20
S127	DRAFT_REQUIREM	10	15	20
S127	DELIVERY_COMPO	5	10	15
S127	PREVENT_FUNDIN	1	6	11
S127	PLACING_RAW_MA	10	15	20
S127	PROGRAM_REQUIR	5	10	15
S127	MODEL_CALIBRAT	15	20	25
S127	WIDGET_MALFUNC	12	15	18
S127	RELIABILITY	8	15	20
S127	COMMUNICATION	7	14	18
S127	TECHNICAL_PERF	12	16	23
S128	FINISH_ON_TIME	11	19	24
S128	ADAPTER_QUALIT	7	13	18
S128	FUNDING_AWARD	7	11	16
S128	CAPABILITY_IMP	16	21	24
S128	DRAFT_REQUIREM	12	17	23
S128	DELIVERY_COMPO	15	20	25
S128	PREVENT_FUNDIN	7	16	22
S128	PLACING_RAW_MA	10	17	22
S128	PROGRAM_REQUIR	12	15	23
S128	MODEL_CALIBRAT	11	16	23
S128	WIDGET_MALFUNC	15	20	25
S128	RELIABILITY	5	10	15
S128	COMMUNICATION	5	10	15
S128	TECHNICAL_PERF	15	18	23
S129	FINISH_ON_TIME	16	20	25
S129	ADAPTER_QUALIT	6	13	20
S129	FUNDING_AWARD	11	18	25
S129	CAPABILITY_IMP	3	7	12
S129	DRAFT_REQUIREM	8	12	18
S129	DELIVERY_COMPO	11	17	23
S129	PREVENT_FUNDIN	3	8	12
S129	PLACING_RAW_MA	6	13	18
S129	PROGRAM_REQUIR	1	5	10
S129	MODEL_CALIBRAT	16	20	25

S129	WIDGET_MALFUNC	7	13	22
S129	RELIABILITY	8	14	17
S129	COMMUNICATION	10	12	15
S129	TECHNICAL_PERF	5	15	18
S130	FINISH_ON_TIME	1	8	13
S130	ADAPTER_QUALIT	1	11	24
S130	FUNDING_AWARD	8	9	12
S130	CAPABILITY_IMP	4	11	15
S130	DRAFT_REQUIREM	18	19	23
S130	DELIVERY_COMPO	5	11	20
S130	PREVENT_FUNDIN	20	22	24
S130	PLACING_RAW_MA	12	16	22
S130	PROGRAM_REQUIR	5	7	10
S130	MODEL_CALIBRAT	3	8	10
S130	WIDGET_MALFUNC	5	9	25
S130	RELIABILITY	4	15	23
S130	COMMUNICATION	11	14	16
S130	TECHNICAL_PERF	10	11	17
S131	FINISH_ON_TIME	4	12	19
S131	ADAPTER_QUALIT	11	19	25
S131	FUNDING_AWARD	3	13	19
S131	CAPABILITY_IMP	7	8	12
S131	DRAFT_REQUIREM	12	21	25
S131	DELIVERY_COMPO	2	23	24
S131	PREVENT_FUNDIN	13	23	25
S131	PLACING_RAW_MA	4	6	21
S131	PROGRAM_REQUIR	2	7	9
S131	MODEL_CALIBRAT	1	16	25
S131	WIDGET_MALFUNC	4	14	22
S131	RELIABILITY	2	19	23
S131	COMMUNICATION	2	4	15
S131	TECHNICAL_PERF	11	14	25
S132	FINISH_ON_TIME	2	11	18
S132	ADAPTER_QUALIT	11	19	20
S132	FUNDING_AWARD	1	11	20
S132	CAPABILITY_IMP	14	20	21
S132	DRAFT_REQUIREM	14	22	25

S132	DELIVERY_COMPO	6	8	16
S132	PREVENT_FUNDIN	3	11	23
S132	PLACING_RAW_MA	3	8	10
S132	PROGRAM_REQUIR	2	9	20
S132	MODEL_CALIBRAT	5	13	21
S132	WIDGET_MALFUNC	11	12	13
S132	RELIABILITY	14	19	24
S132	COMMUNICATION	6	11	23
S132	TECHNICAL_PERF	8	15	23
PWDM	FINISH_ON_TIME	11	19	24
PWDM	ADAPTER_QUALIT	7	13	18
PWDM	FUNDING_AWARD	7	11	16
PWDM	CAPABILITY_IMP	16	21	24
PWDM	DRAFT_REQUIREM	12	17	23
PWDM	DELIVERY_COMPO	15	20	25
PWDM	PREVENT_FUNDIN	7	16	22
PWDM	PLACING_RAW_MA	10	17	22
PWDM	PROGRAM_REQUIR	12	15	23
PWDM	MODEL_CALIBRAT	11	16	23
PWDM	WIDGET_MALFUNC	15	20	25
PWDM	RELIABILITY	5	10	15
PWDM	COMMUNICATION	5	10	15
PWDM	TECHNICAL_PERF	15	18	23

Appendix E: WinBUGS Scripts

The following are the model scripts used for the Bayesian inference to determine the optima Bayes Factor and Bayesian Information Criteria. The first is the exponential model depicted in Figure E - 1 script.

H4M1a

```
model { #//Dana & Kelly p.47, 103 and 106 Using CVM, BIC, BAYES FACTOR
# MODEL = EXPONENTIAL
# WIDGET = CODE A
  for (i in 1:N) {
    time[i] ~ dexp(lambda)
    time.rep[i] ~ dexp(lambda)

    time.ranked[i] <- ranked(time[], i)
    time.rep.ranked[i] <- ranked(time.rep[], i)

    F.obs[i] <- cumulative(time[i], time.ranked[i])
    F.rep[i] <- cumulative(time.rep[i], time.rep.ranked[i])

    diff.obs[i] <- pow(F.obs[i] - (2*i-1)/(2*N),2)
    diff.rep[i] <- pow(F.rep[i] - (2*i-1)/(2*N),2)

    log.like[i] <- log(lambda) - lambda * time[i]
  }

fi  lambda ~ dgamma(alpha,beta)

  CVM.obs <- sum(diff.obs[])
  CVM.rep <- sum(diff.rep[])

  log.like.tot <- sum(log.like[])
  BIC <- -2 * log.like.tot + log(N)

  p.value <- step(CVM.rep - CVM.obs) #p.value should ~ 0.5
                                     #//Dana & Kelly p.36
```

Figure E - 1 – Exponential OpenBUGS Script

It is followed by the Weibull model depicted in Figure E - 2 script.

H4M2a

```
model { #//Dana & Kelly p.47, 103 and 106 Using CVM, BIC, BAYES FACTOR
# MODEL = WEIBULL
# WIDGET = CODE A
  for (i in 1:N) {
    time[i] ~ dweib(beta,lambda)
    time.rep[i] ~ dweib(beta,lambda)

    time.ranked[i] <- ranked(time[], i)
    time.rep.ranked[i] <- ranked(time.rep[], i)

    F.obs[i] <- cumulative(time[i], time.ranked[i])
    F.rep[i] <- cumulative(time.rep[i], time.rep.ranked[i])

    diff.obs[i] <- pow(F.obs[i] - (2*i-1)/(2*N),2)
    diff.rep[i] <- pow(F.rep[i] - (2*i-1)/(2*N),2)

    log.like[i] <- log(lambda) + log(beta) + (beta - 1) *
    log(time[i]) - lambda * pow(time[i], beta)
  }

  lambda ~ dgamma(0.0001, 0.0001)
  beta ~ dgamma(0.0001, 0.0001)

  CVM.obs <- sum(diff.obs[])
  CVM.rep <- sum(diff.rep[])

  log.like.tot <- sum(log.like[])
  BIC <- -2 * log.like.tot + log(N)

  p.value <- step(CVM.rep - CVM.obs) #p.value should ~ 0.5
                                     #//Dana & Kelly p.36

}
```

Figure E - 2 – Weibull OpenBUGS Script

Lastly, the last model considered is the Lognormal model depicted in Figure E -3 script.

H4M3a

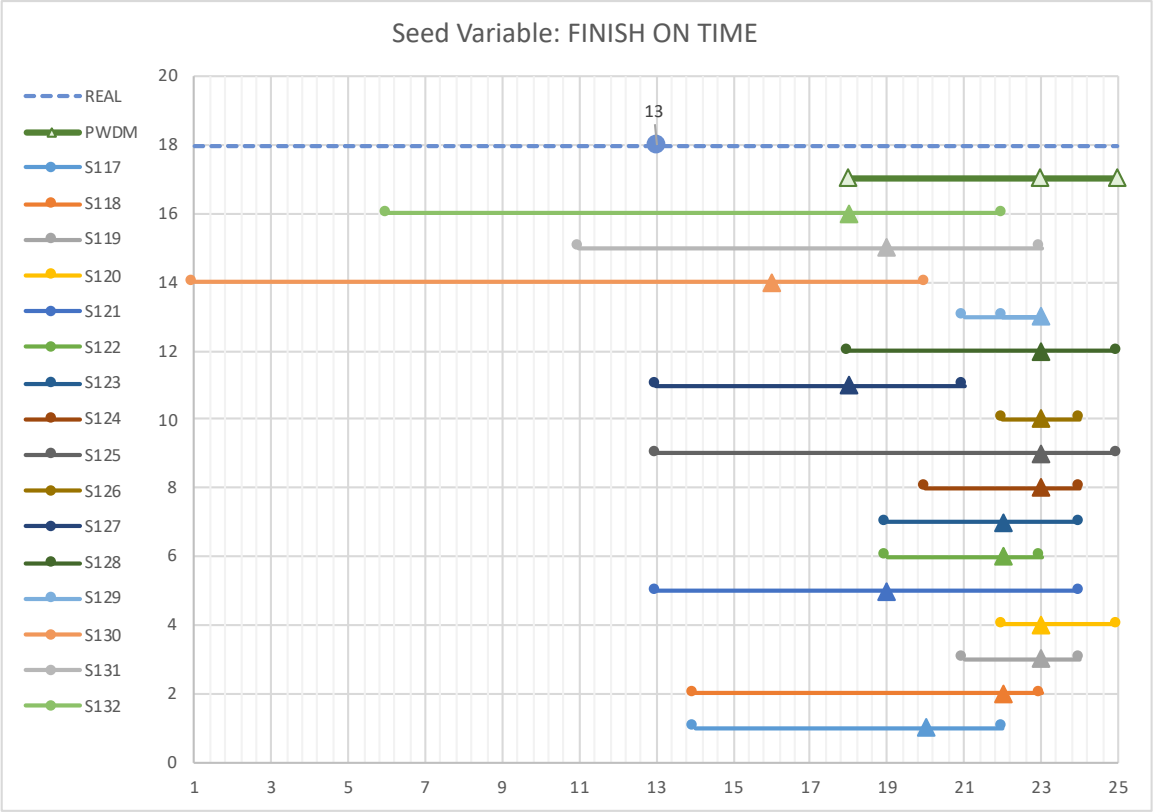
```
model { #//Dana & Kelly p.47, 103 and 106 Using CVM, BIC, BAYES FACTOR
# MODEL = LOGNORMAL
# WIDGET = CODE A
  for (i in 1:N) {
    time[i] ~ dlnorm(mu, tau)
    time.rep[i] ~ dlnorm(mu, tau)

    time.ranked[i] <- ranked(time[], i)
    time.rep.ranked[i] <- ranked(time.rep[], i)

    F.obs[i] <- cumulative(time[i], time.ranked[i])
    F.rep[i] <- cumulative(time.rep[i], time.rep.ranked[i])

    diff.obs[i] <- pow(F.obs[i] - (2*i-1)/(2*N),2)
    diff.rep[i] <- pow(F.rep[i] - (2*i-1)/(2*N),2)

    log.like[i] <- -0.5 * (log(2) + log(3.14159))
    - log(sigma) - log(time[i]) - pow(log(time[i]) - mu, 2)
    / (2 * pow(sigma,2))
  }

  mu ~ dflat()
  sigma ~ dgamma(0.0001, 0.0001)
  tau <- pow(sigma, -2)

  CVM.obs <- sum(diff.obs[])
  CVM.rep <- sum(diff.rep[])

  log.like.tot <- sum(log.like[])
  BIC <- -2 * log.like.tot + log(N)

  p.value <- step(CVM.rep - CVM.obs) #p.value should ~ 0.5
                                     #//Dana & Kelly p.36

}
```

Figure E - 3 – Lognormal OpenBUGS Script

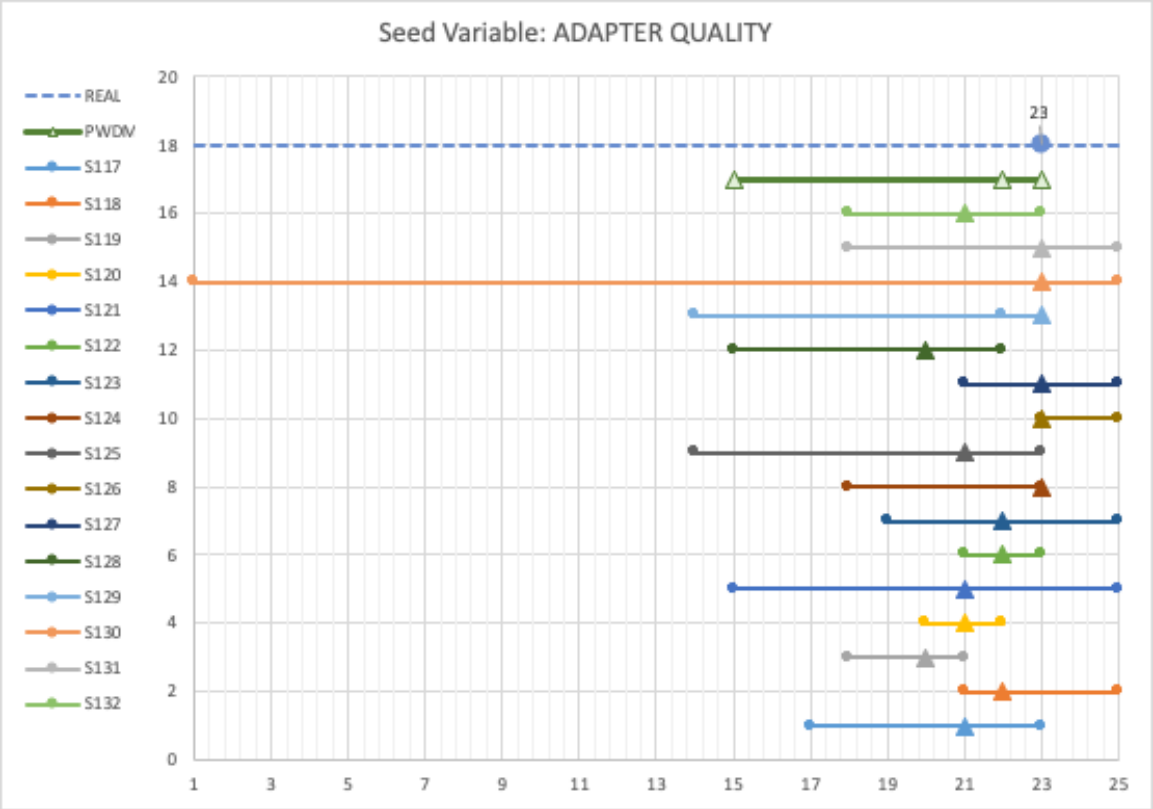
Once one of the above scripts is loaded with the data, the process to start the numerical solution via the MCMC. Instructions on how to run a model in OpenBUGS are outside the scope of this research study, but instructions can be quickly found in the software tutorial. The MCMC numerical approximations to the BI are displayed by the software, based on the script and data loaded.

Appendix F: Range Graphs

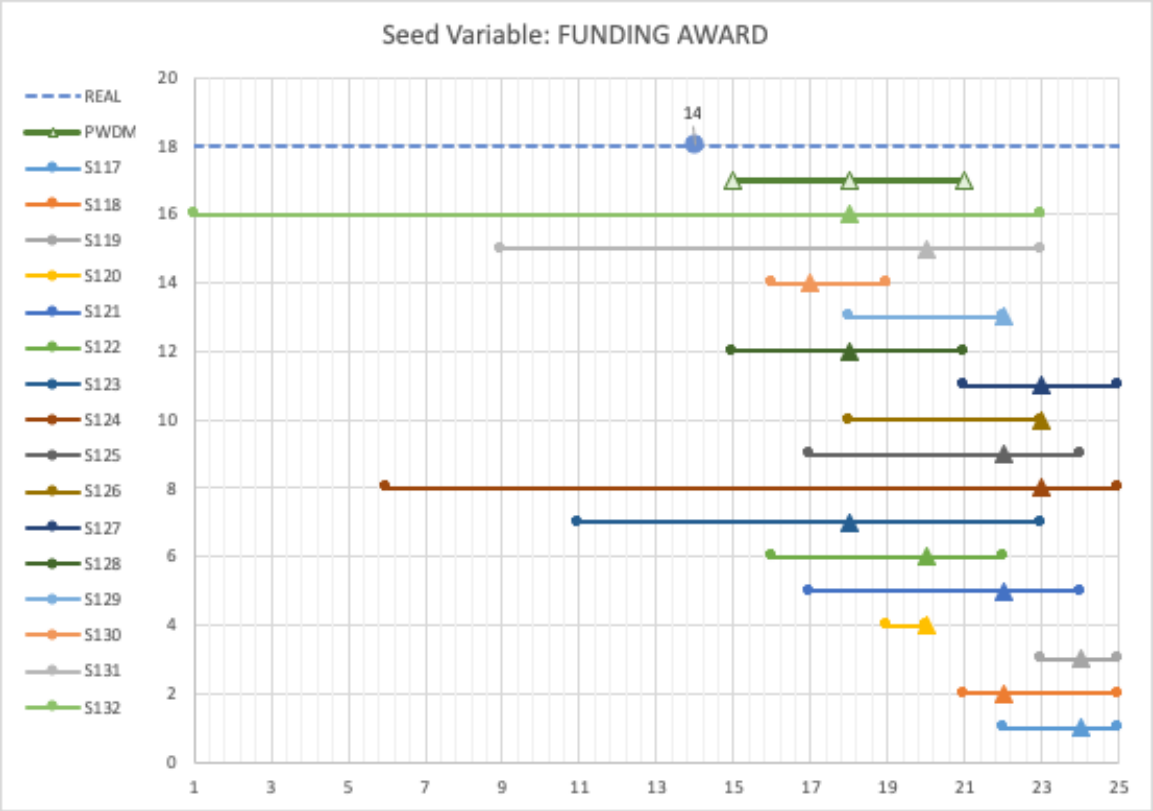
The following are the range graphs not shown in Chapter 4 for the seed variables. The definitions of each seed variable is explained in Table 10.



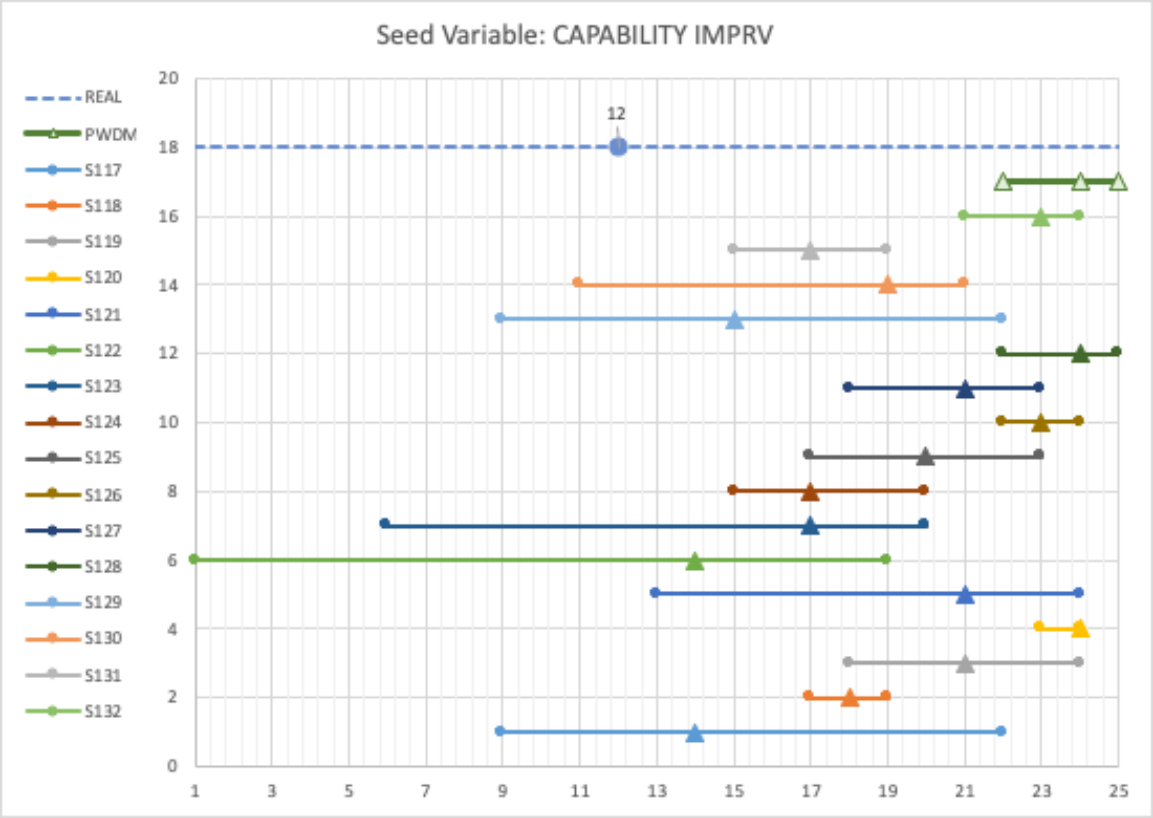
F 1 - Finish on Time Seed Variable



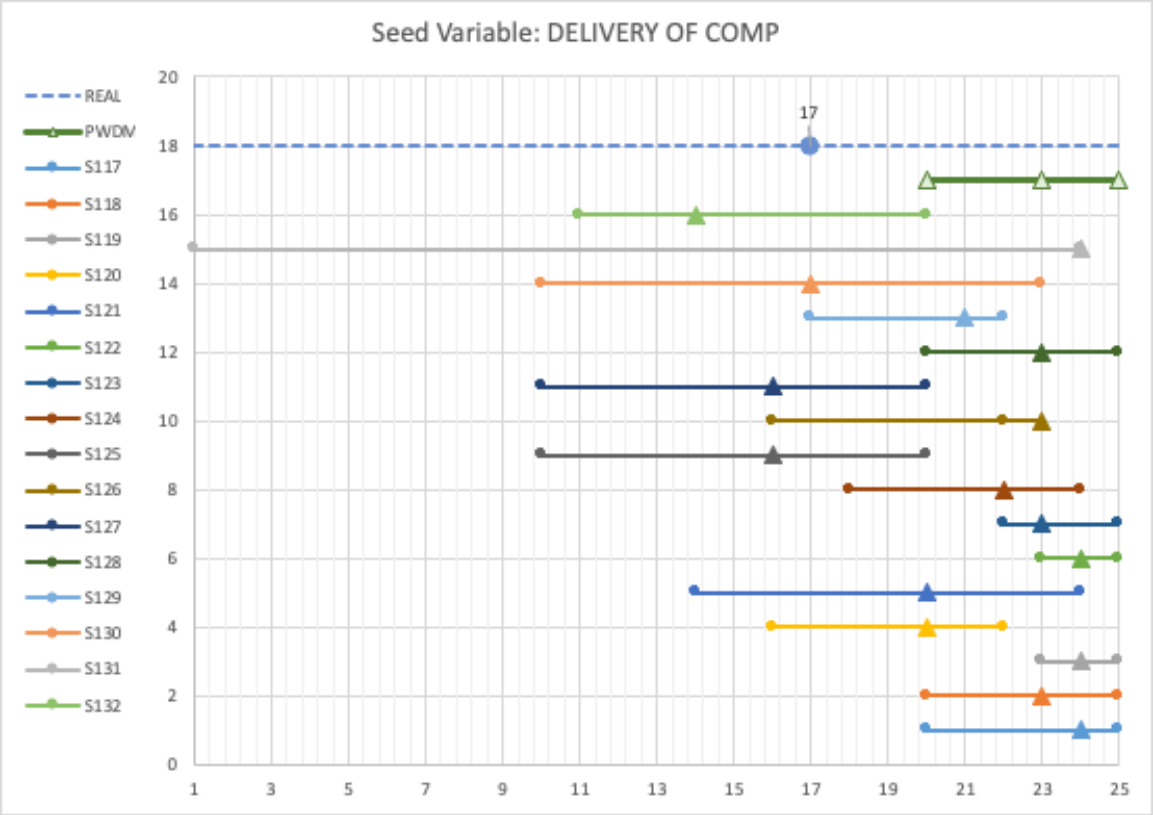
F 2 - Adapter Quality Seed Variable



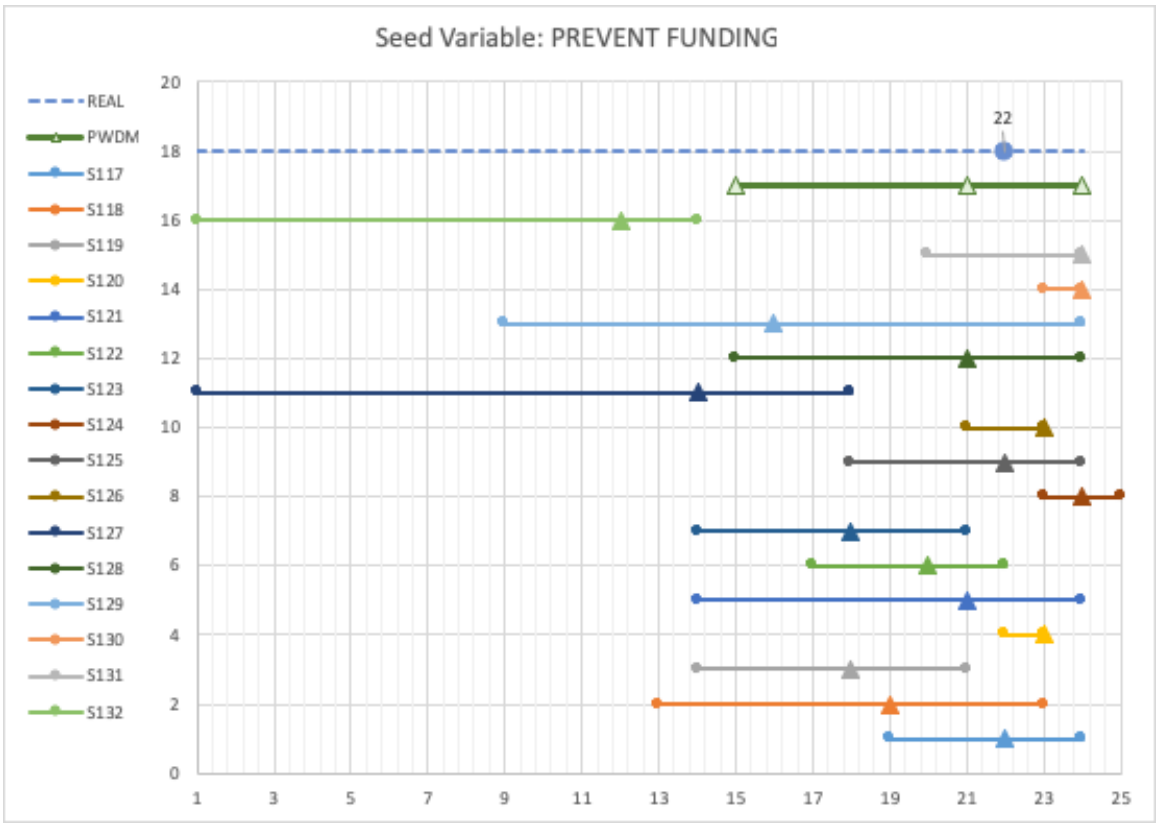
F 3 – Funding Award Seed Variable



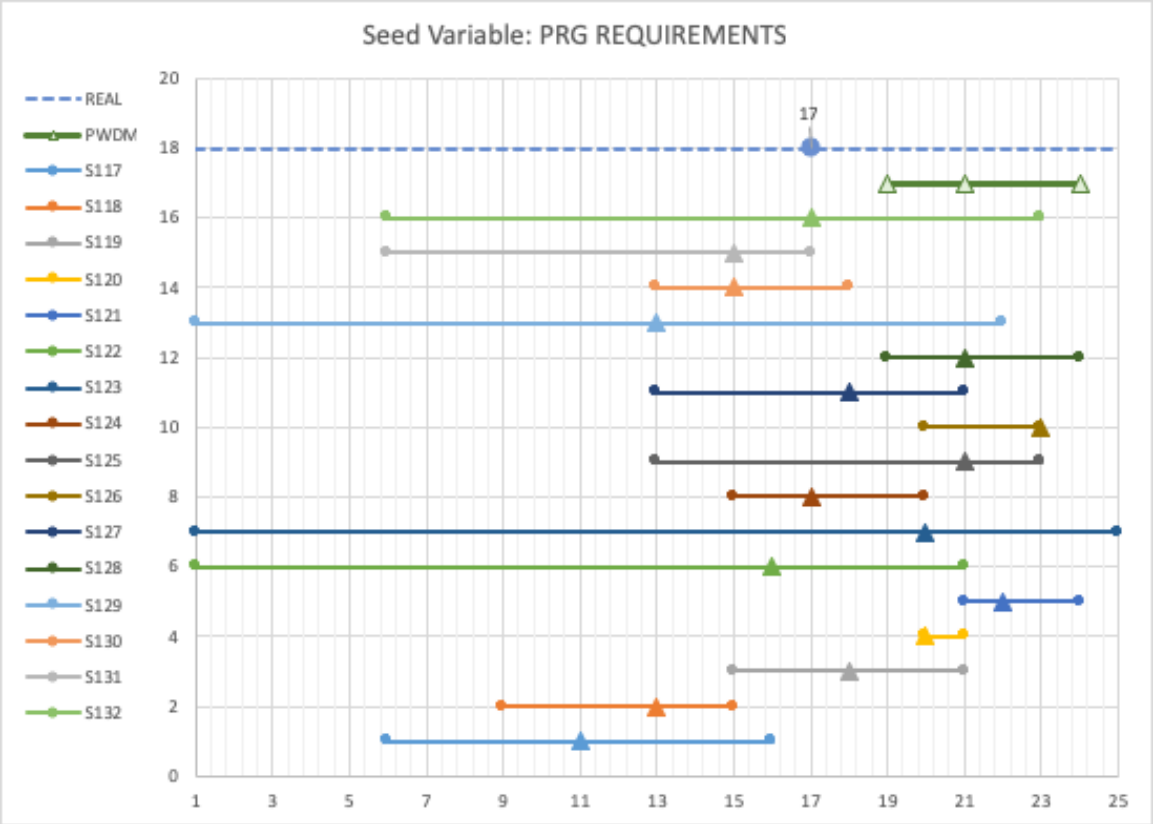
F 4 - Capability Improvement Seed Variable



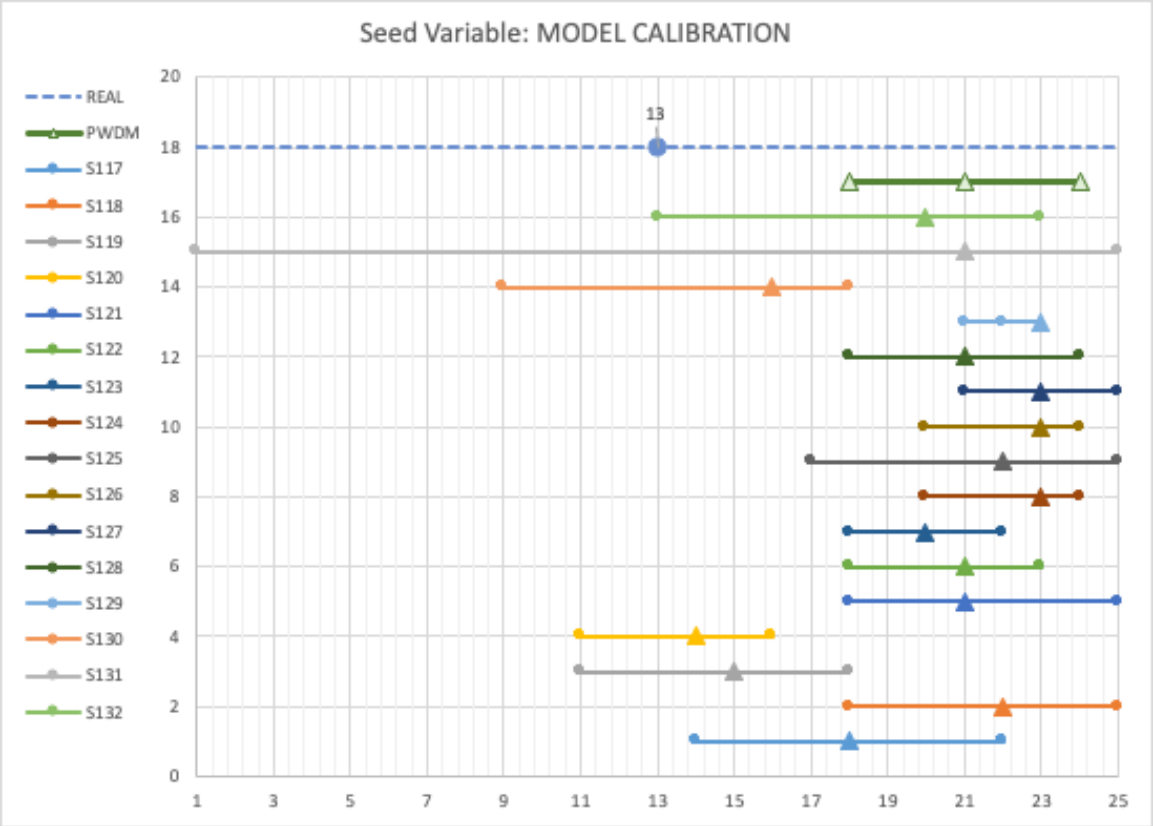
F 5 - Delivery of Component Seed Variable



F 6 - Prevent Funding Issues Seed Variable



F 7 - Program Requirements Seed Variable



F 8 - Model Calibration