3-2020

The Utility of Self-Assessment in Predicting Program Office Estimate Accuracy

Dana P. Luketic

THE UTILITY OF SELF-ASSESSMENT IN PREDICTING PROGRAM OFFICE ESTIMATE ACCURACY

THESIS

Dana P. Luketic, Master Sergeant, USAF

AFIT-ENV-MS-20-M-225

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENV-MS-20-M-225

THE UTILITY OF SELF-ASSESSMENT IN PREDICTING PROGRAM OFFICE ESTIMATE ACCURACY

THESIS

Presented to the Faculty

Department of Systems Engineering and Management

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cost Analysis

Dana P. Luketic, BS

Master Sergeant, USAF

March 2020

AFIT-ENV-MS-20-M-225

THE UTILITY OF SELF-ASSESSMENT IN PREDICTING PROGRAM OFFICE ESTIMATE ACCURACY

Dana P. Luketic, BS
Master Sergeant, USAF

Committee Membership:

Lt Col S. T. Drylie, PhD
Chair

Dr. J. D. Ritschel
Member

Dr. R. D. Fass
Member

Abstract

The ability of the Program Offices to provide accurate cost estimates is an essential element in planning and programming. Historically, cost estimating has led to budget overruns and continues to be an area of scrutiny and concern. A series of legislative reforms have sought to address each of these perceived underlying causes which are located at all levels of decision making – from the SPO to CADE. The current study is specifically interested in determining how well SPOs are doing. There have not been comprehensive studies on SPO performance. In large part, this deficiency is due to the inability to systematically assess the SPOs. However, a new consolidation of data by AFLCMC has recently made it possible to do such a study. The AFLCMC's program office estimates in this study will look at the SPOs of AFLCMC and evaluate their cost estimates for growth and determine if their established method of self-assessment provides a predictor of the overall future accuracy of the program estimate.

Acknowledgments

I would first like to thank my research advisor, Lt Col Scott Drylie, for his valuable guidance, support, and mentorship throughout this process. I truly enjoyed working with you and I could not have completed this milestone without you. I would also like to sincerely thank my thesis committee for their insight, constructive criticism and encouragement. Lastly, I would like to thank my classmates, friends, and family for being incredibly supportive.

Dana P. Luketic

Table of Contents

Page

List of Figures

Page

List of Tables

THE UTILITY OF SELF-ASSESSMENT IN PREDICTING PROGRAM OFFICE ESTIMATE ACCURACY

I. Introduction

Background

The ability of the Program Offices to provide accurate cost estimates is an essential element in planning and programming. Historically, cost estimating has led to budget overruns and continues to be an area of scrutiny and concern. A series of legislative reforms have sought to address each of these perceived underlying causes which are located at all levels of decision making – from the System Program Office (SPO) to Cost Assessment and Data Enterprise (CADE). From the 1983 Nunn-McCurdy Act, and continuing with the Federal Acquisition Streamlining Act of 1994, the Federal Acquisition Reform Act of 1996, as well as the 2009 Weapon Systems Acquisition Reform Act (WSARA) legislation has tried to reduce cost growth (Schwartz, 2016). Even with reform, cost growth continues to be an issue within the Air Force. The issues may involve numerous agents within the complex corporate structure of estimating, as well as planning and programming. How well SPOs are doing, and the quality and value of their self-assessment tools is the specific subject of the current study. There has been no comprehensive studies on SPO performance. The lack of research is due, in large part, to the inability to systematically assess SPOs. However, with the support of Air Force Life Cycle Management Center (AFLCMC) providing access to their historical records and database, it is now possible to do such a study.

Problem Statement

A common conclusion is that cost growth has remained an issue despite intensive efforts at reform. A 2007 RAND study noted that in the three decades prior cost growth remained high, with no significant improvement (Younossi, 2007). More recently, the GAO concluded that most programs continue to proceed without the key knowledge of technologies, design, cost, and schedule, essential to good acquisition outcomes (GAO, 2018).

It is not clear, however, how well SPOs, in particular, are doing. The indictments above alight upon a variety of agents, but not necessarily the SPOs. The reason is that the program baseline is established through a corporate process. And that process involves higher authorities who ultimately choose the baseline. Those authorities for the Air Force process is the Air Force Cost Analysis Agency (AFCAA) and then the Office of the Secretary of Defense, Cost Assessment and Program Evaluation (OSD-CAPE). Their involvement is one of developing their own independent estimates, and reaching a single one based on the multiple perspectives. The GAO goes so far as to say that the most robust study, and thus most credible and important, is the one conducted by AFCAA (GAO, 2009). It is important to note that in the traditional documentation and the ensuing literature, the SPO's initial estimate is generally shrouded and lost – it accuracy left to be an unknown for systematic review.

The current study intends to address two primary gaps identified in literature on the current state of cost estimating by those working with Air Force SPOs. First, using the AFLCMC's internal program office estimates, which have been retained in PowerPoint slide shows used in program reviews, this study can finally systematically

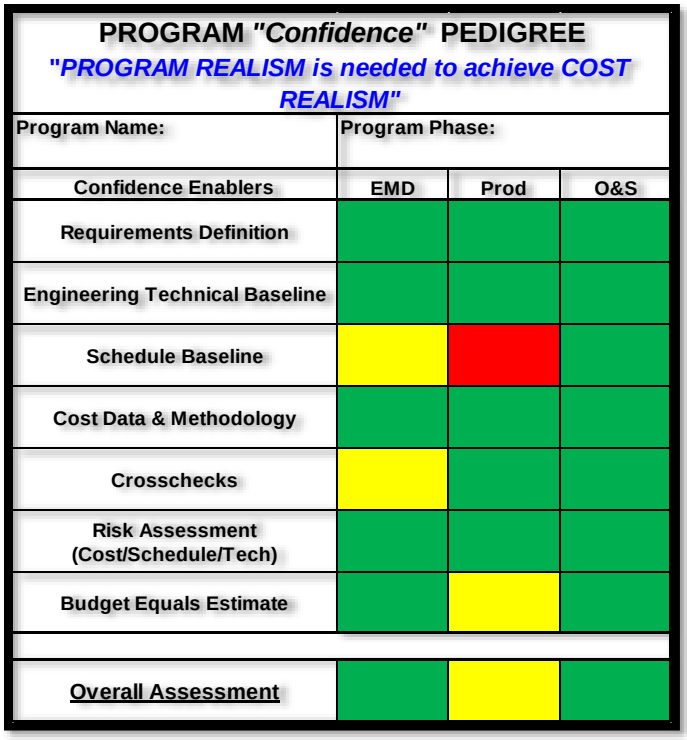
determine the accuracy of SPO. AFLCMC has provided 480 program offices estimates spanning sixteen year of historical data for research. The hope is this data will aid in determining the accuracy provides an important commentary on the SPOs' value in the process. Much is made of the need for independent estimates to account for potential bias within the SPO. Namely, the SPO could be too invested in a program to identify problems. Or, the SPO could be too close to the contractors to properly scrutinize their actions. But the personnel in SPOs have easy access to the same publications that make these claims as anyone else. Most likely, these publications were part of their own training and education. It would seem, then, that through a self-awareness, internal checks might overcome biases, and professionalism trump cronyism. In the end, one must put the question of accuracy to the test. Until now, there has been no way to systematically do so.

Second, this study can provide a new way to address the common sentiment contained within the GAO's claim: DoD lacks key knowledge for good estimating. Such a claim is uncontested in a general sense. Cost estimators recognize that their knowledge is imperfect. Creating new technologies, aiming for new metrics, and employing new engineering process all involve uncertainties. One would like a better-defined end goal, or better historical analogies, or perfectly prescient subject matter experts. But lacking these, cost estimators as a profession deal with uncertainty, and they do so in systematic ways. Numerous handbooks and guides set the methods for quantifying uncertainty, and these methods involve, most commonly, the utilization of variance parameters. The eventual baseline, then, represents merely a "most likely" point within a range of possible outcomes. But a novel question would be to ask how they internalize this deficiency of

knowledge and communicate it in an internal business process. That is to say that a "most likely" estimate is a best one can produce, but whether that is the 50th percentile of the 70th using the esoteric language of statistics, what is the general sentiment in common parlance, and discourse? The PowerPoint slides retained by AFLCMC allow one to glance behind the numbers and begin to understand the relationship between knowledge and confidence in a different way, one that perhaps inputs the people, the culture, and the business environment back into the study of DoD and its long term reform initiatives.

SPO documentation shows that SPOs have employed an internal device, an internal review process, since 2003 with the aim of qualitatively assessing the end product. This review is a business process tool. Such tools have been studied within business literature, but have received scant attention in DoD acquisitions literature. The question for this second study, then, will be to determine if such a tool has worth within DoD, namely by accurately predicting future problems and successes.

Figure 1 shows a representative self-assessment tool that has been in place at AFLCMC since 2003, and which will be at the center of this second study. The figure is a Program Confidence Pedigree. Although there are minor variations of this pedigree format, they all have in common the scale of confidence of Green, Yellow, and Red. Each SPO has available the descriptions of what these colors represent. (Appendices A to C). This scale is used to assess seven different "confidence enabler" categories that comprise the pedigree. Together these seven enablers permit an Overall Assessment.



PROGRAM *"Confidence"* PEDIGREE *"PROGRAM REALISM is needed to achieve COST REALISM"*			
Program Name:	Program Phase:		
Confidence Enablers	EMD	Prod	O&S
Requirements Definition			
Engineering Technical Baseline			
Schedule Baseline			
Cost Data & Methodology			
Crosschecks			
Risk Assessment (Cost/Schedule/Tech)			
Budget Equals Estimate			
Overall Assessment			

Figure 1: Program Pedigree Confidence Self-Assessment

This study of self-assessment data could lead to further answering the problems of cost growth as well as giving insight to program offices' ability to assess themselves. An important part of reform is the establishment of repeatable systems for internal controls. While AFLCMC has employed this internal control, it remains to be determined whether its personnel have employed it in a manner that would head off problems. It is credible to think that something so summary might lead to shrouded inputs with derogatory colors of yellow and red being underrepresented.

Research Objectives/Questions/Hypotheses

The self-assessment tool contains numerous elements, each yielding potentially important information. It is a rich source of data that has yet to be mined. The following

questions represent what the study deems to be the most relevant for a first assessment of the tool.

1. Does confidence in a program increase through the course of a program, as one would expect if the tool has some general validity?

2. What is the value of program office overall self-assessments as predictors of cost estimating accuracy?

3. Which of the self-assessment confidence enablers are most predictive of future variance?

4. When does a program office self-assessment peak in confidence during a program? When does confidence stabilize in a program?

5. To what extent can we attribute changes in program estimates to factors controllable by cost estimators?

6. Have cost estimates improved over time within the SPO? And has the technique of self-assessment proved more valuable over time?

Methodology

The AFLCMC database was used for data collection and variables that are used in the analysis. The data was analyzed through descriptive statistics, contingency tables and Ordinary Least Squares (OLS) regression models. Descriptive statistics will describe the basic features of the dataset. Contingency tables are used to set up the relationship between the program office assessments and different cost growth or cost variance thresholds and the regression model will estimate the relationship between a dependent variable the independent variables.

Research Contribution

The predominant focus of this research is to provide the Program Office data on their self-assessments and look at how they aligned with their estimates to see if they have improved over time. The results of the research provide AFLCMC and other DoD entities potential insight into the how Program offices view their methods and potentially provide avenues for better cost estimates. Results of the study may not apply to all acquisition program types, but may provide key information for program offices to improve cost estimates.

II. Literature Review

Chapter Overview

This chapter first discusses the Program Office's Estimates (POE) and their estimation techniques. Then it looks at Cost Growth studies and research, factors of cost growth, and then cost growth modeling. Lastly, the chapter reviews the self-assessment literature, looking at the value it brings to organizations, issues with self-assessment, and how it can be fully utilized in program office cost estimation.

Program Office Estimates

POEs are detailed estimates of acquisition costs normally required for high-level decisions. The POE is a full Life Cycle Cost Estimate (LCCE) which is initially built by the program office staff early in the program's life (GAO, 2009). Estimates are then produced on a recurring basis which are tied to major program reviews to include Milestones A, B, and C or to acquire funding for program changes throughout its lifespan (GAO, 2009).

The POEs serve numerous purposes. The initial program estimate serves as the base point for all subsequent tracking and auditing purposes (DAU, 2019). Further, the POE is used at acquisition program milestones and decision reviews to assess whether the system's cost is affordable and/or consistent for long-range investment and force structure plans. These estimates are also the basis for budget requests to Congress as a vehicle for inputs to the programming and budgeting phases of the Planning, Programming, Budgeting and Execution process (PPBE) (DAU, 2019).

POEs are mandated and governed at the Department of Defense (DoD) level. The policies and procedures for the preparation of POEs for Major Defense Acquisition Programs (MDAPs) and Major Automated Information System (MAIS) programs at key acquisition events, as well as the requirements for cost data collection, are described in the Department of Defense Instruction (DoDI) 5000.73 (DAU, 2019). This regulation mandates that the POE be presented to OSD Cost Assessment and Program Evaluation (CAPE) prior to milestones required to create an Independent Cost Estimates (ICE) for all Acquisition Category (ACAT) I and IA programs.

Program Offices utilize multiple methods to come up with their estimates. They use the Work Breakdown Structure (WBS) as a framework for cost estimation, program planning and reporting. The WBS defines the total system to be developed or produced (AcqNotes, 2019). There are 4 Cost Estimating Categories for LCC: Research and Development, Investment, Operating and Support, and Disposal (DoD 5000.4) The LCCs are divided into the five major Appropriations Categories. Research, Development Test and Evaluation (RDT&E), Procurement; Operations and Maintenance (O&M), Military Construction (MILCON), Military Personnel (MILPERS) (AcqNotes, 2019) Funding is received in the respective cost category under applicable appropriation.

There are a number of cost estimating techniques that can be used to develop a program office estimate. The main techniques are: Analogy, Engineering Estimate, Parametric, and Actual Cost. The first method discussed is an analogy. It is often used when a program is very early in its life cycle and it does not have a detailed breakdown or actuals to build its estimate. It often uses historical actuals from an analogous program for its estimate (GAO, 2009). Analogy assumes similar characteristics for the new

weapons systems to historical weapon systems and can use complexity factors that account for new technology from subject matter experts (SME). Its benefit is that it is quick to develop but on the downside, they are subjective epically when there is no historical weapon system that is a direct match to a new weapon system (GAO, 2009)

Another method is engineering build-up. It relies on the detail of the weapon system's WBS. The more details, the more accurate the estimate will be. It starts at the lowest WBS level and adds or builds from there. Each level is added to the next until all components and levels are calculated. While it can identify cost drivers, it can be very time consuming and very reliant on good data. If the data is not available at the lowest level then engineering build-up estimates becomes very difficult (GAO, 2009).

Parametric cost estimating uses statistical methods such as regression to develop a Cost Estimating Relationship (CER). This method uses the CER to predict the future costs of a new weapon system based on the historical data. Regression is a common method used in developing a CER because it allows the estimator to make statistical inferences that are important to consider when using parametric estimating (GAO, 2009). Finally, actual costs use trends from prototypes or early production items to project estimates of future costs for the same system. These projections may be made at various levels of detail, depending on the availability of data, without the data actuals cannot be calculated.

There are other methods and the viability of these methods that will be discussed further as the cost growth literature looks at how useful and accurate cost estimating methods are. The discussed cost estimation methods are utilized by the program offices

for estimates to try and limit cost growth. The literature on cost growth cites the estimation methods among some of the known causes for growth.

Cost estimation is described as a science and an art. The ever-increasing complexity of technology, software density, system integration complexity, and the like make estimating a total system's development cost, at the inception of major development activities, an increasingly challenging endeavor (Younossi, 2007).
If the strength of initial program estimating accuracy cannot improve, then continued monitoring and tracking of the cost growth is required. Next, with the knowledge of a program office and how they get their estimates, the literature will dive into cost growth and how it is assessed or categorized.

Cost Growth

Cost growth is an ongoing DOD issue. From the literature, cost growth is defined as "the difference between the initial estimate of the total acquisition cost for a program and the most recent or final estimate adjusted for inflation and quantity changes" (Jarvaise et. al, 1996). Others define cost growth as the tendency of the unit cost of a system to increase during the course of the acquisition process (Singer, 1982). The impact of cost growth is that it compromises the ability of DoD to procure the total number and type of weapon systems needed to meet mission requirements. For this research Jarvaise's definition is in line with our data collected.

Now that cost growth has been defined, there have been numerous RAND and AFIT studies on cost growth. The RAND Corporation is a nonprofit research organization providing objective analysis and solutions to challenges facing the public

and private sectors. Early studies include Drezner who studied Cost Growth Factors (CGF) in weapon systems. For 128 Major Defense Acquisition Programs (MDAPs) Drezner found a 20% cost growth from the initial baseline estimate through the life cycle of the program (Drezner, 1993). Additionally, the GAO Analysis of DOD Major Defense Acquisition Program 2008 Portfolios stated that change from the first estimate was 26% cost growth (GAO, 2009). However, other cost growth studies said that this number is too conservative. IDA and RAND studies looked to better capture the full extent of cost growth over program's entire lifecycle. They showed average adjusted total cost growth for the completed program is 46% from MS II and 16% from MS III, about 20% higher growth than the previously studied (Arena, 2006). With the magnitude of cost growth now understood the review focuses on the reasons or factors leading to cost growth.

The factors responsible for cost growth have been identified in several ways. In early studies factors where identified. Since then they have been built upon by more recent research. Table 1 Searle's research developed factors for cost growth specifically: Planning Difficulties, Risk Elements, and Management Inefficiencies.

Table 1: Factors Affecting Cost Growth (Searle, 1999)

Planning Difficulties	Risk Elements	Management Inefficiencies
1) Incomplete Definition of Work	6) Unforeseeable Conditions	11) Disorganized Work Direction and Productivity
2) Interface Incompatibilities	7) Unpredictable Regulatory Funding delays	12) Subcontracting
3) Changes; Failure to Anticipate Needs.	8) Unforeseen Technical Difficulties	13) Unnecessary Work or "Gold Plating"
4) Estimating Uncertainties; Poor Estimating	9) Uncontrollable Forces	14) Project Control
5) Optimistic Assumptions	10) Unanticipated Economic Conditions	15) Work Load Projections

These show, first, that planning difficulties tend to prevent realistic and early estimates of the final cost of a program. Second, risk elements are the factors inherent to the system and are neither controllable nor predictable. Finally, management inefficiencies are the factors that are considered as controllable by the management. He found planning difficulties and management inefficiencies only played a minor role in growth with much of the growth in uncontrollable risk elements (Searle, 1999). His findings suggest that a significant amount of cost growth is outside of estimation and due to uncontrollable factors for the SPO -- an important idea when related to what extent can we attribute changes in program estimates to controllable factors.

In another early study, Calcutt identified five factors: requirements definition, cost estimating, program management, contracting, and budgetary. "Requirements Definition" refers to poor initial requirement definition, poor performance/cost trade-offs during development, and changes in quantity requirements. "Cost Estimating" refers to errors due to limitations of cost estimating procedures, poor inflation estimates, top down pressure to reduce estimates, and a lack of valid independent cost estimates. "Program Management" refers to the lack of program management expertise, mismanagement/human error, over optimism, and schedule concurrency. "Contracting" refers to the lack of competition, use of wrong type of contract, inconsistent contract management/administrative procedures, and too much contractor oversight and reporting requirements. (5) "Budgetary" refers to funding instabilities within DoD, funding instabilities caused by congressional decisions, and inefficient production rates due to

stretching out programs. This approach facilitates correlating the factors attributed to cost growth with the initiatives DoD has taken to reduce cost growth. (Calcutt, 1993).

More recent studies have shown different cost growth factors. Deneve focusing on the issues with DoD cost estimating instead of post-estimation drivers. He focused on what they were estimating and not the not the accuracy of the cost estimate. He found four categorical variables with strong relationships to cost growth factors: program type, iteration, funding years, and the number of services. The Cost Growth Factors (CGF) were predictors of the total cost from the baseline estimate with 71 % of initial estimates are made better by at least correcting in the direction indicated by the predicted CGF (Deneve, 2015).

Arena found three major categories affecting cost growth: schedule factors, acquisition strategy, and other factors. The study identified schedule slip and program duration as the factors that affects cost growth with total program CGFs. Although there was mixed evidence of the effectiveness of acquisition strategies, the study found cost growth due to decisions outside of the control of program managers increased program costs(Arena, 2006).

Finally, Bolten looked at SARs, finding cost growth or variance tied to the following categories: quantity, schedule, engineering, estimating, economic, other, and support. These variance categories along with how they classified the variance data showed the "realized" cost growth. The cost growth was found to be in four major categories: (1) errors in estimation and planning, (2) decisions by the government, (3) financial matters, and (4) miscellaneous sources. Of the four major categories, decisions by the government dominate the overall growth in both development and procurement.

For development, decisions account for 31% of the 57% cost growth; for procurement, they account for 57% of the 75% cost growth (Bolten, 2008). The factors of discussed in the literature above are similar to what the POE database from our research uses as categories when explaining reasons for change in cost estimates and relates to research question five.

These similar factors give credence to the SPOs using them as a bases for identifying and categorizing cost growth. It is important to have a standardized methodology in this categorization. Proper cost growth identification can help identify problem areas and differences among programs.

Cost growth literature has shown the magnitude and factors of it. Table 2 provides a list of the AFIT studies, specifically, in recent years.

Table 2: AFIT Cost Growth Research

Moore & White	2003	A Regression Approach for Estimating Procurement Cost
White et al.	2004	Using Logistic and Multiple Regression to Estimate Engineering Cost
Lucas	2004	Creating Cost Growth Models for the Engineering and Manufacturing Development Phase of Acquisition Using Logistic and Multiple Regression
Rossetti & White	2004	A Two-Pronged Approach to Estimate Procurement Cost Growth in Major DoD Weapon Systems
Bielecki & White	2005	Estimating Cost Growth From Schedule Changes: A Regression Approach
Genest & White	2005	Predicting RDT&E Cost Growth
Monaco & White	2005	Extending Cost Growth Estimation to Predict Schedule Risk
McDaniel & White	2007	Predicting Engineering and Schedule Procurement Cost Growth for Major DoD Programs
Foreman	2007	Predicting the Effect of Longitudinal Variables on Cost and Schedule Performance
Rusnock	2008	Predicting Cost and Schedule Growth For Military and Civil Space Systems
Brown et al.	2015	Time Phasing Aircraft R&D Using the Weibull and Beta Distributions
Kozlak et. al	2016	Predicting Cost Growth Using Programs Reviews and Milestones for DoD Aircraft
D'Amico	2017	A Longitudinal Study and Color Rating System of Acquisition Cost Growth

AFIT studies have used historical data and regression analysis to estimate cost growth. First, White et al. developed a seven-variable cost growth model with funding variables, time variable, and length of program predicting cost growth. Others followed the early works of White et al. utilizing this research in predicting cost growth (White et al., 2004). Bielecki et al. (2005) and Moore et al. (2005) generated models to predict cost growth in different funding appropriations using regression analysis. Moore et al. (2005)

modeled procurement cost growth during the EMD phase. Bielecki et al. (2005) modeled RDT&E budget cost growth during the EMD phase of the program lifecycle.

Further studies by Lucas (2004), Rosetti et al. (2004), Genest et al. (2005), McDaniel et al., (2007) furthered the ability to predict cost growth in DoD weapon systems. Lucas (2004) focused on developing a model to predict a range or variance of cost growth. Rossetti et al. (2004) modeled procurement and support costs during the EMD phase. Genest et al., (2005) modeled the pre EMD phase of a weapon system to predict cost growth. Monaco et al. (2005) modeled the predicted schedule risk. McDaniel et al., (2007) modeled cost risk early in EMD looking reduce cost growth. Foreman (2007) furthered Monaco et al predicting schedule slip and using this to predict cost growth. Rusnock (2008) modeled cost estimators in predicting schedule growth in space systems. Deneve et al. (2015) as discussed in the cost growth factors previously modeled more realistic cost estimates using factors for cost growth while at AFIT.

Brown et al. (2015) studied development expenditures at 50% program completion. He compared commonly cited 60/40 "rule of thumb," which assumes 60% expenditures at 50% schedule. Finding the estimation of budgets by 6.5% higher, on average 60/40 model. Kozlak et al. (2016) modeled cost growth factors as predictors of cost growth at the four reviews finding that at 6.5 years after MS-B a program sustains about 91% of the total program cost growth. D'Amico et al. (2017) modeled a color rating matrix for cost growth factors finding RDT&E having the biggest impact. His color rating mirror the self-assessment inputs of this study, however the cost growth outputs this study not similar.

Self-Assessment

The program office estimates have conducted self-assessments since 2003. The SPOs however, have not utilized the self-assessment as a tool observe the estimates. The specific measures could identify causes in lower overall confidence in the program and potentially greater variance in cost estimation. Self-assessment have long been studied to determine their validity as a tool for improvement. The literature shows self-assessments can be an organizational improvement tool increasing effectiveness and efficiency through introspection. Additionally, it has found multiple key points to assessing value. Such tools could be a useful method for cost estimation improvement within program offices and provide a tool for cost estimators and managers.

A self-assessment tool is valuable because it has shown that organizations can track policies and programs over focused areas and identify actions to strengthen procedures and improve performance (Keyser et. al, 2008). By using self-assessment tools, an organization can determine how to solve problems and react to pressure. Self-Assessment needs to have organizational buy-in and clear guidance for the outcomes properly implemented.

Self-assessments need some sort of framework or structure when being developed. This can be achieved through the integration of key requirements. Stecher studied the framework for organizational self-assessments. It contained seven criteria that form the basis for organizational self-assessments: Leadership, Strategic planning, Customer and market focus, Information and analysis, Human resource focus, Process management, and Business results. (Stecher et. al, 2004). These criteria are their pillars for the basis for organizational self-assessments.

Organizations must also be able to provide exceptional service within an appropriate budget. “Performance is increasingly judged by the efficiency of the organization by the cost per service, the number of outputs per employee, the number of outputs per person per year, or the average value of grants per person” (Lusthaus et. al, 1999). No matter how large an organization is, it viewed by the value they provide both quantitative and qualitatively.

The literature has established the importance of self-assessment criteria being identified and the proper framework established. The research studied in this thesis utilized the categories from The Program “Confidence” Pedigree Self-Assessment and Overall Confidence previously seen in Figure 1 as the framework for self-assessment.

Organizations often conduct self-assessments, rate performance and manage strategic issues, with an end goal of improved performance. Organizational self-assessment can be a useful tool to implement change or adjust planning to improve performance. “Self-assessment is based on a detailed organizational profile and a strategic plan linked to clearly identified goals and reinforced by an information and analysis system to collect data and monitor progress toward those goals” (Stecher et. al, 2004). Effectiveness of reaching those goals can be difficult to assess, a clearly defined mission statement is needed for the organization to assess effectiveness.

Lusthaus identified the shift from Assessment to ‘Self-Assessment’ in an organization relies on a model to guide the assessment of the strengths and weaknesses of a targeted organization. For organizational assessment to realize its potential, assessors need to engage in a teacher-learner type of relationships with the source requiring the

information as well as the one providing it. Collecting and analyze specific objective data, and then 'teach' managers about their organizations. (Lusthaus et. al, 1999)

A direct participatory approaches by the technical expertise of the evaluator with perspectives from inside the organization allows for the best gathering of information. "This self-assessment process not only teaches the members of an organization how to collect and analyze data by themselves, it guides them while making decisions based on it, drawing conclusions, and generating solutions" (Lusthaus et. al, 1999) Self-assessment can be defined as a 'learning process' some organizations may support a brainstorming stage for more information before making a decision, others need technical support with tools and instruments to guide them. Lusthaus studies are extensive in the process of how to properly build a self-assessment. It is up to the organization correctly build and utilize this effective tool and helps answer research question one when determining the value of self-assessment as a predictor.

Improving effectiveness and efficiency are primary reasons for self-assessments. Bartuseviciene's research focused on utilizing self-assessment for in these two areas. They utilized a mixed-method approach to determine if an organization had the prerequisites to learn from the evaluation. Then a real-time study of the organization and self-evaluation were implemented. The results of Bartuseviciene indicated the need for supportive organizational contexts, structures, and processes of evaluation used for learning is to occur throughout the organization. (Bartuseviciene, 2013) Feedback shows that self-assessment has a positive impact. The act of engaging in a reflective process about one's work, including focusing on outcomes and evaluation should ideally not just be an afterthought to programming.

However, the participants of the study were not sure how the self-assessment study could apply to their daily work and voiced concerns for further content clarification as well as continuous implementation support if self-assessment become an integrated practice. (Bartuseviciene, 2013) The study's intent was to aid managers striving for better performance results.

The research found that effectiveness and efficiency are performance measures that organizations can use to assess their performance. "Efficiency is oriented towards successful input transformation into outputs, where effectiveness measures how outputs interact with the economic and social environment" (Bartuseviciene, 2013). Self-assessment can be applied to team of workers or overall organizations to help identify strengths and weaknesses and improve performance. With proper feedback as well as evaluation for supervisors, peers, and subordinates, employees evaluate their own performance and participate in setting goals. The research supported self-assessment is a valuable and useful tool for organizations, provided that it is properly administered and reviewed.

Efforts have been done develop models for organizational self-assessments. Siow developed a multiple-attribute decision-making (MADM) modeling framework and methodology (Siow, 2018). The model scored self-assessments and focused on initial development and application for management. It can use thousands of criteria and several alternatives in the assessment. The model's applications are varied including project management and organizational self-assessments. It was proven valid and reliable when applied to real data (Soiw, 2018) the research in this study uses the program office overall

confidence self-assessment to try and model cost growth predictors in contrast to the specific models of self-assessment just discussed.

Not all research supports self-assessment. Dunning studied the empirical evidence on flaws in self-assessment. Research into a wide variety of domains examined how accurately people judge themselves. Finding, "first a fairly small correlations between personal perceptions and objective performance. Second, people tend to be too optimistic about skill, expertise, and future prospects" (Dunning, 2004). Kolar studied self-assessment vs outside assessment. Results slightly favored the predictive validity of judgement made by single acquaintance over self-judgment. "It significantly favored the aggregate judgment of two acquaintances over self-judgment. These findings imply that the most valid source judgment may not be self-assessment but the consensus of multiple peer assessors" (Kolar, 1996).

Harris completed a meta-analysis on the reviews of self–supervisor, self–peer, and peer–supervisor ratings. They measured the mean correlation values of the three groups with the closer the ρ was to 1 the higher level of correlation. "The results indicated a relatively high correlation between peer and supervisor ratings (ρ= .62) but only a moderate correlation between self-supervisor (ρ= .35) and self-peer ratings (ρ= .36)" (Harris, 1988). The analysis looks at how the supervisor's, peers, and the assessor's views are different. The results show peer and supervisor assessment were closer related than one's own self-assessment when compared. This questions who should do the assessment.

The literature shows there is value in self-assessment, but this is limited by the methods used. Self-assessment models have been be made to produce better assessments,

but these can be limited as assessor tend to over judge their self-performance. The literature supports supervisor and peer reviews of self-assessment and this can be noted as a best practice for the program office overall confidence self-assessments.

Summary

There is little literature on cost estimation *within* program offices, but there is a vast amount on cost growth, estimates and factors for mitigation. The literature also shows that in spite of regulation, there is still little system-wide improvement in cost estimating. Additionally, the program office only has a role to play in estimates and many factors are beyond their controls. Self-assessment can be a valuable tool for organizations as it may help them improve effectiveness and efficiency. In order to be implemented, it needs to have organizational buy-in, clear guidance, and the outcomes properly understood. Chapter III goes into methodology of the data collection and analysis for the self-assessment of program office estimates.

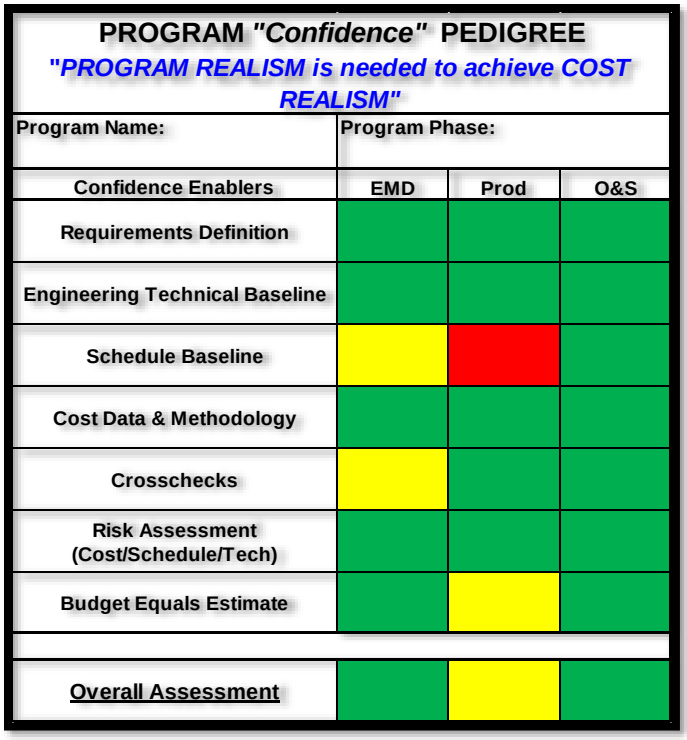
III. Methodology

Chapter Overview

The first part of this chapter covers the data source and database, and then goes into how the data was collected and variables that are used in the analysis. Then the chapter covers the methodology for the contingency tables and Fisher's Exact Tests. Contingency tables are used to set up the relationship between the program office assessments and different cost growth or cost variance thresholds, while the Fisher's Exact Tests test for statistical significance.

Data Source/Database Summary

The data for this research comes from Air Force Life Cycle Management (AFLCMC) Program Office Estimates (POE) briefing slides. They were used to retrieve all of the self-assessment data from the Program Pedigree Confidence slide Figure 2 below. The data was then recorded within a database for all the measure as well as overall confidence. The set of programs range from ACAT I to ACAT III.



PROGRAM *"Confidence"* PEDIGREE *"PROGRAM REALISM is needed to achieve COST REALISM"*			
Program Name:	Program Phase:		
Confidence Enablers	EMD	Prod	O&S
Requirements Definition			
Engineering Technical Baseline			
Schedule Baseline			
Cost Data & Methodology			
Crosschecks			
Risk Assessment (Cost/Schedule/Tech)			
Budget Equals Estimate			
Overall Assessment			

Figure 2: Program Pedigree Confidence Self-Assessment

The dataset was initially comprised of 480 POEs across 162 programs. Some these programs did have not all the necessary data for an analysis. Only 445 of the 480 POEs complied with standard estimate submission. There were several parts of the POE submission missing or they only completed the budget portion of the POE submission. Next, there were only 308 that completed the self-assessment tables. Some removed the tables, others chose to leave them unfilled or blank. Finally, the data was analyzed against the cost estimation data. An additional 34 did not have POE change estimation data. The final dataset had 274 POE for statistical analysis. The POEs provided development and production data. See Table 3 below:

Table 3: Database Inclusions/Exclusions

Criteria	Δ POEs	Total POEs	Δ Programs	Total Programs
Initial POE data provided by AFLCMC	+ 480	480	+ 201	201
POE did not meet typical submission formatting	- 35	445	- 10	191
POE missing self-assessment data	- 137	308	- 12	179
POE missing change estimation data	- 34	274	- 17	162

While the data source is a rich set of programmatic insights, which the current study can only begin to mine, it poses some challenges as well. A particular challenge for the current study is that the POEs do not recur in a common regular basis. Some programs may have one estimate per year, for others, every few years or irregularly, as needed. How then to normalize the data in the quest to systematically determine a POEs ability to predict future costs? One must identify how complete a program is when the first POE appears, and one must determine how complete a program is for each subsequent POE. And then one can begin to compare programs with like starting points and like intervals.

The method used, was to take the first POE then the deviation was measured by measuring the percent difference between the First POE and Last POEs. As stated they have different intervals; thus to normalize the data, the cost growth was averaged per annum for the period between the POE submissions. Equation:

$$\text{Percent Cost Growth Per Year} = \frac{(\text{POE Obligations First Year} - \text{POE Obligations Latest})}{((\text{POE Date First Year} - \text{POE Date Latest})/12\text{Months})}$$

Table 4 below provides a breakdown of the percent cost growth per annum data defining the elements of calculation above.

Table 4: Percent Cost Growth per Annum Data

Term	Definition
POE Date – First Year	The date of the initial or first Program Office Estimate Submitted
POE Date – Latest	The date of the latest or final Program Office Estimate Submitted
POE Obligations–First Year	The expenses reported from the initial or first Program Office Estimate Submitted
POE Obligations-Latest	The expenses reported from the latest or final Program Office Estimate Submitted
POE Date Delta	The differences in month of the date reported from the first to latest Program Office Estimate Submitted
POE Obligations Delta	The difference in expenses reported from the first to latest Program Office Estimate Submitted
Percent Cost Growth Per Year	The difference in expenses reported from the first to latest Program Office Estimate Submitted divided by the difference in dates of Program Office Estimate Submitted calculated per annum.

Additionally the percent complete was calculated by taking the expensed portion of the POEs divided by the total cost. This was done as a measure of the program's life cycle using the percent work completed as stages: early, middle, late middle, and near complete. The self-assessment analysis for program progress was shown by percent work complete. Equation:

Percent Work Completed = POE Obligations Earliest/POE Total Cost

Data Collection

Collecting the data was a manual process of coding POE data from individual files provided by AFLCMC. It included 480 POEs spanning from 2003 to 2018. From the Program "Confidence" Pedigree self-assessment inputs. Variable data was collected under seven designators: requirements definition, engineering technical baseline, schedule baseline, cost data, crosschecks, risk assessment, and budget equals estimate. The seven designators lead to an overall confidence self-assessment variable that rated the entire program office estimate submission. Data was collected for the development and production. The data was collected for use in this research and provided AFLCMC with self-assessments that can be studied and analyzed to potentially aid in future POE submissions.

Additionally, further variables were created for analysis of programs for assessment and estimation. The appropriation growth per annum (cost growth), and the POE percent change variables were created. The cost growth has the weakness of only being a normalized average over the life span of the POE and is a limitation of the dataset. This is due the data coming from the AFLCMC database where they have the cost growth between two POE submission and the timeframe but due to all programs having different durations the data is averaged per annum so different programs can be compared.

Dummy variables were created to analyze POEs with high variance. Two notional thresholds were chosen for the category of high variance: +/- 10% and +/- 20%. There is no empirical foundation for such thresholds. However, these thresholds are significant in an institutional sense. Nunn-McCurdy oversight legislation identifies a

"significant" breach is when the current baseline estimate is breached by 15 % (Arena, 2014) 10% was, therefore, chosen as initial marker, low enough to stay below the "significant breach" but still high enough to raise concern with the cost estimate.

The Percent Completed variable determined the percent complete by taking the cost of the work complete and dividing it by the total cost. Percent complete was binned into 4 groups of 25% intervals from 0% to 100% with 25% quartile breaks. The 25% bin included data with percent complete from 0% to 25%, while the 50% bin included data from the 25% to 50%, etc. Again, this simplifying approach was necessary because of the irregularity of the POE frequency. For example, the Electronic Board Operation Support System had a POE interval of less than a year, from 26 Sept 2013 to 11 Sept 2014. Conversely the MQ-9 Reaper an unmanned aerial vehicle (UAV) had a POE interval of greater than eight years with dates from 2 Nov 2009 to 14 Mar 2018. These are two extremes, but they accentuate the differences in the interval submissions.

An additional work completed variable was generated to provide greater granularity in the tables. Percent complete was binned into 10 groups this time with 10% intervals from 0% to 100%. The first 10% bin included data with percent complete from 0% to 10%, while the 20% bin included data from the 10% to 20%, and so on. Table 5 shows data base variables utilized in the statistical analysis.

Table 5: Database Variables

Self-Assessment/Cost Estimation Variables for Descriptive Statistics	Dependent Variables for Contingency Tables	Independent Variables for Contingency Tables
Appropriation Growth Per Annum	Overall Confidence	POE Iteration
Percentage Change in Estimate	Overall Confidence G	Appropriation Growth Per Annum
Overall Confidence	Overall Confidence Y	Requirement definition
POE Year	Overall Confidence R	Engineering technical baseline
DV Percent Change >10%	Percentage Change in Estimate	Schedule baseline
DV Percent Change >20%	Reason for Change in Estimate	Cost data
DV Percent Work Complete 25%	DV Percent Change >10%	Crosscheck
DV Percent Work Complete 10%	DV Percent Change >20%	Risk assessment
	DV Percent Work Complete 25%	Budget equals estimate
	DV Percent Work Complete 10%	

Descriptive Statistical Analysis

Utilizing the database of 162 programs, self-assessment and cost growth estimates from POEs was analyzed through descriptive statistics. The descriptive statistics consisted of the mean, standard deviation, maximum, minimum, first quartile, median, and third quartile. Descriptive statistical analysis is used to summarize the self-assessment and cost growth data in the database. The mean and median values show the typical self-assessment cost growth data for cost estimates. The standard deviation and

variance are quantitative descriptions of the dispersion in the data. The minimum and maximum values show the entire range of values in the data. Finally, quartiles provide a visualization of the distribution of values. The quartiles are measure the 25th percentile and 75th percentile of the data.

Contingency Table Analysis

The dataset generated many categorical variables to study. The Program Pedigree Confidence Self-Assessment generated the categorical variables of engineering technical baseline, schedule baseline, etc. Categorical variables were established for each measurement of self-assessment in the data collection. The database contingency table analysis identified potential variables effecting overall confidence in self-assessment as well as cost growth in SPO programs. Continuous variables of cost growth in percentages were looked at as categorical binary variables, or dummy variables (+/-10%, +/- 20%). Additional dummy variables tying them to the color codes for green, red, and yellow self-assessment levels.

Contingency tables are used to test whether or not independent variables can predict dependent variables. The typical statistical test that is used for contingency tables is the Pearson test. However, the Pearson test fails requires a large sample size for the p-value provided. Fisher's Exact Tests are geared to account for small sample sizes within contingency tables. Fisher's Exact Test is different than the other statistical tests because it is unconditioned in the number of rows and columns from its second assumption. The Fisher's Exact Test calculates the probability of getting the observed data. The p-value

determines if there is significant effect. This research utilized a significant threshold p-value of 0.05 to potentially explain the results of the self-assessment data.

In this research the contingency tables are used to test whether or not any of the independent variables of quantity, schedule, engineering, estimating, economic, other, and support could predict the dependent variable cost growth greater than +/- 10% as method to determine what reasons drive large cost growth variations. Additionally, contingency tables are used to test whether the percent work complete is a predictor of Self-assessment overall confidence. Percent work complete was coded for each in 10% intervals to give and idea where the program stands as it progresses through major program reviews to including Milestones A, B, and C, etc. throughout the life cycle of the program. This would determine if the overall confidence improves over the program's life cycle. The tests will be conducted on both development and production phases.

One-way analysis of variance (ANVOA)

A one-way analysis of variance (ANOVA) was used to test the significance self-assessment variables. It looked that the impact of time on the self-assessment overall confidence if the p-value from the ANOVAs is less than 0.05 than the tested variable was considered to be a significant predictor cost growth/cost variance as is relates to the number of the POE submission, a proxy for time in the programs life cycle. ANOVA was also completed on cost growth/cost variance versus percent work complete as second check. The dataset was broken into subsets where the overall confidence Green and overall confidence Yellow were studied separately. Due to the smaller sample size of

some of the smaller subsets the nonparametric Wilcoxon and Kruskal-Wallace Tests were used to test whether the differences in the means is significant using an alpha of 0.05.

Summary

This chapter started off with the source of the data how it was collected and categorized. Then the chapter discussed the details of database formulation. Next, went into the independent variables that are used in the analysis. Additionally, the statistical and contingency table analysis were defined along with the thresholds for those estimate factors. After this the chapter went through the methodology of the analysis and the Fisher's Exact Tests. Finally the chapter covered the ANOVA and fit tests.

IV. Analysis and Results

Chapter Overview

The purpose of this chapter is to provide the analysis and results of the methodology outlined in Chapter III. The chapter finds the results of Investigative Questions using descriptive statistical analysis and well as contingency table analysis as measures of self-assessment and cost growth predictors. The results highlight significant p-values for the tests and outcomes of these tests. The chapter concludes with a summary of analysis of the results.

Investigative Questions Answered

4.1 Does confidence in a program increase through the course of a program, as one would expect if the tool has some general validity?

To answer the research question, contingency table analysis was used to look at the program office life cycle. Using the binned variables for percent work complete, the overall confidence levels of Green, Red, and Yellow were compared to the percent work complete with contingency tables. The variable created four 25% quartile breaks of the percent work complete into four bins as explained in Chapter III. The results show the actual occurrences for Green are far less in the first 25% than expected based on the assumption that there is no relationship. Table 10 shows the dataset had only had 94 occurrences in the first quartile -- less than the expected value of 137.98 occurrences. By the final quartile >75% complete Green had 85 occurrences but there was an expected value of only 55.64 occurrences. The p-value for the Fisher's Exact Test was statically significant at <.0001. See Table 6.

Table 6: Overall Confidence vs Work Complete

Contingency Table

OverallConfidence	DV_WK	1	2	3	4	Total
	Count					
	Total %					
	Col %					
	Row %					
	Expected					
G		94	57	60	85	296
		17.67	10.71	11.28	15.98	55.64
		37.90	61.29	65.93	85.00	
		31.76	19.26	20.27	28.72	
		137.985	51.7444	50.6316	55.6391	
R		16	1	1	0	18
		3.01	0.19	0.19	0.00	3.38
		6.45	1.08	1.10	0.00	
		88.89	5.56	5.56	0.00	
		8.39098	3.14662	3.07895	3.38346	
Y		138	35	30	15	218
		25.94	6.58	5.64	2.82	40.98
		55.65	37.63	32.97	15.00	
		63.30	16.06	13.76	6.88	
		101.624	38.109	37.2895	40.9774	
Total		248	93	91	100	532
		46.62	17.48	17.11	18.80	

Tests

N	DF	-LogLike	RSquare (U)
532	6	41.339954	0.0964

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	82.680	<.0001*
Pearson	76.101	<.0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	2.07e-22	<.0001*

The Figure 3 below gives a nice visual respentation of the contingency as well. In the first 25% quartile, green was underrepresented and yellow over represented when compaired to their expected values, by the last quartile the opposite was true. There is a statically significant difference in the number of programs with an overall confidence of green as the percent work completed increases or as the POE submissions increase. Figure 5 clearly depicts that the greater the percent work completed, the higher the occurance of green overall confidence. It is iteresting to note that only two of the eighteen Red incidences occur after 25% work complete.Once this level of work completion is achieved, programs do not remain Red. The overrepesentation of Green at a statistically significant level shows there is a dependent relationship between overall confidence and

the percent work complete. The significance of the results comports with program office self-assessments becoming more optimistic over the course a program life cycle. The finding seems logical as the SPOs' familiarity, methodology, confidence in the program should only improve over time.

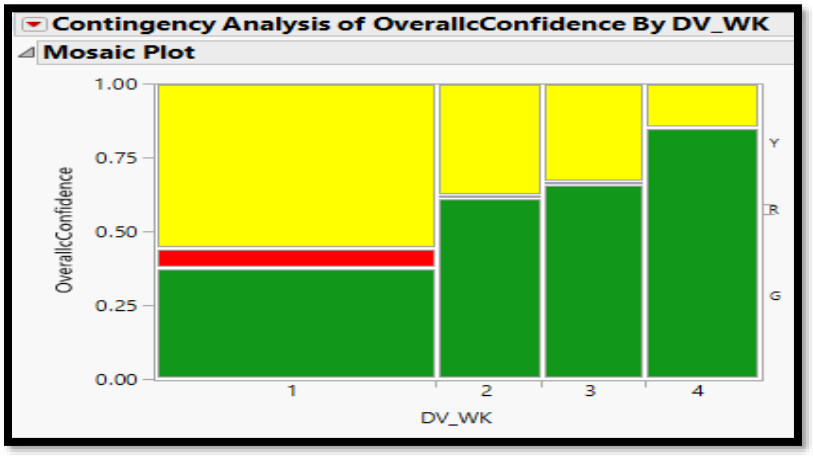
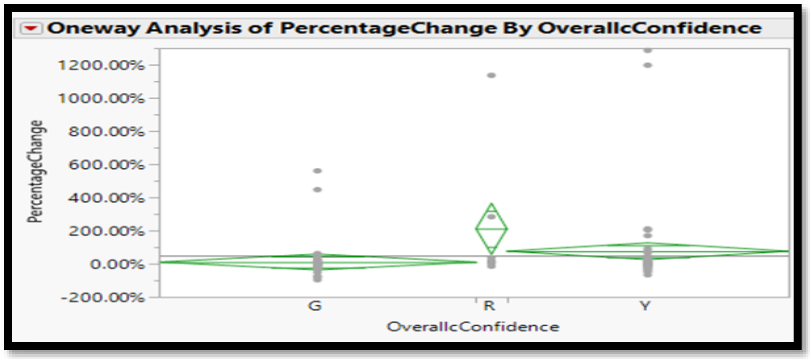


Figure 3: Overall Confidence vs Percent Work Complete-25% Intervals

4.2 What is the value of program office overall self-assessments as predictors of cost estimating accuracy?

To answer research question two a one-way analysis of variance (ANOVA) was performed on the data performed. The ANOVA bins each program based on its confidence category (Green, Red, and Yellow). It then uses nonparametric tests to compare the medians of the estimate deviations that follow.



Oneway Anova

Summary of Fit

Rsquare	0.053431
Adj Rsquare	0.039909
Root Mean Square Error	2.059031
Mean of Response	0.497329
Observations (or Sum Wgts)	143

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
OverallcConfidence	2	33.50384	16.7519	3.9513	0.0214*
Error	140	593.54516	4.2396		
C. Total	142	627.04900			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
G	72	0.10782	0.24266	-0.3719	0.5876
R	7	2.10939	0.77824	0.5708	3.6480
Y	64	0.75921	0.25738	0.2504	1.2681

Std Error uses a pooled estimate of error variance

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
G	72	4298.50	5184.00	59.701	-3.573
R	7	701.000	504.000	100.143	1.838
Y	64	5296.50	4608.00	82.758	2.793

1-Way Test, ChiSquare Approximation

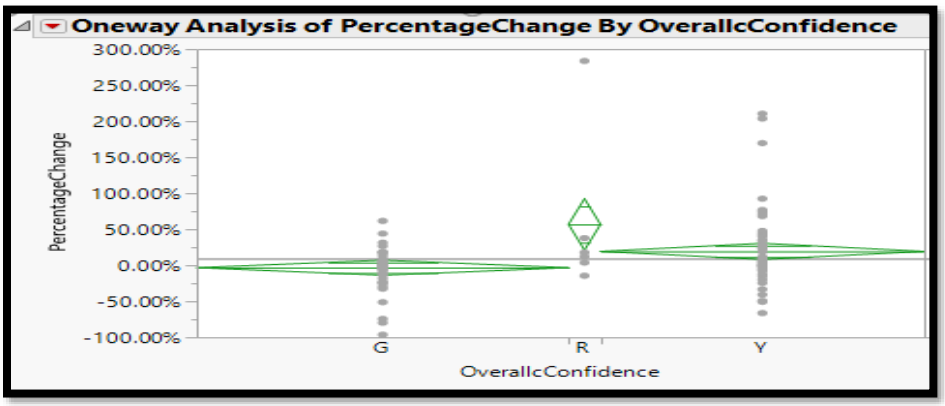
ChiSquare	DF	Prob>ChiSq
13.8939	2	0.0010*

Figure 4: One-way ANOVA Overall Confidence vs Percent Change in Cost Estimate

The ANOVA results in figure 4 show Overall Confidence rating significantly corresponds with the ensuing the Percent Change in Estimates. The means for each category are 10.7% for the Green, 210.9% for Red and 75.9% for the Yellow. The f-statistic produced a p-value of 0.02. Due to the ordinal data and small sets of some of the

samples, the nonparametric Wilcoxon/Kruskal-Wallis Tests were used to compare the medians and confirm that these differences were statistically significant. The test results had a p-value of .001 confirming the statistical significance

Following this first examination, five significant outliers were removed to determine whether the means were still significantly different after their removal. The results remained significant in this analysis as well Figure 5 below. The analysis of the data showed means to be -3.3% for the Green, 56.8% for Red and 19.4% for the Yellow. This analysis supported the initial finding and assures that the difference in means in the initial test was not due to the outliers but the means were different at a statically significant level. Again Wilcoxon/Kruskal-Wallis Tests confirmed the these medians were statistically significant with a p-value of .001



Oneway Anova

Summary of Fit

Rsquare	0.105209
Adj Rsquare	0.091854
Root Mean Square Error	0.442182
Mean of Response	0.094186
Observations (or Sum Wgts)	137

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
OverallcConfidence	2	3.080620	1.54031	7.8778	0.0006*
Error	134	26.200369	0.19553		
C. Total	136	29.280989			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
G	70	-0.03302	0.05285	-0.1375	0.07151
R	6	0.56772	0.18052	0.2107	0.92475
Y	61	0.19358	0.05662	0.0816	0.30555

Std Error uses a pooled estimate of error variance

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
G	70	4021.50	4830.00	57.4500	-3.479
R	6	561.000	414.000	93.5000	1.541
Y	61	4870.50	4209.00	79.8443	2.863

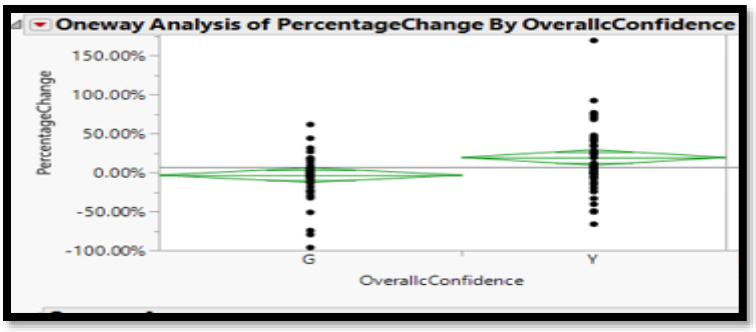
1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
12.7666	2	0.0017*

Figure 5: One-way ANOVA Overall Confidence vs Percent Change without outliers

After confirming difference in means from the first test, the removal of the Red submissions was considered, due to the small number of Red values. A third ANOVA test was run removing the Red inputs, and the data was analyzed again, this time only

comparing the Green and Yellow means. The results remained significant with p-value of .001. Additionally, the F Ratio increased to 10.86 further confirming that Green and Yellow means are significantly different, as a greater F Ratio means greater variation among group means Figure 6 below. The Wilcoxon/Kruskal-Wallis Tests confirmed the these differences as well comparing the medians and resulting in a p-value of .001



Oneway Anova

Summary of Fit

Rsquare	0.077658
Adj Rsquare	0.070508
Root Mean Square Error	0.392543
Mean of Response	0.072498
Observations (or Sum Wgts)	131

t Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
OverallcConfidence	1	1.673613	1.67361	10.8613	0.0013*
Error	129	19.877573	0.15409		
C. Total	130	21.551186			

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
G	70	-0.03302	0.04692	-0.1258	0.05981
Y	61	0.19358	0.05026	0.0941	0.29302

Std Error uses a pooled estimate of error variance

Wilcoxon / Kruskal-Wallis Tests (Rank Sums)

Level	Count	Score Sum	Expected Score	Score Mean	(Mean-Mean0)/Std0
G	70	3919.50	4620.00	55.9929	-3.230
Y	61	4726.50	4026.00	77.4836	3.230

2-Sample Test, Normal Approximation

S	Z	Prob>|Z|
4726.5	3.22994	0.0012*

1-Way Test, ChiSquare Approximation

ChiSquare	DF	Prob>ChiSq
10.4474	1	0.0012*

Figure 6: One-way ANOVA Overall Confidence vs Percent Change excluding Red

Additionally, when doing the initial scatterplot review of the dataset, a boxplot was produced (Figure 7). It analyzed the Green, Red, Yellow overall confidence rating. Green has a much smaller inner quartile range of -5.39% to 3.68% when compared to Yellow's inner quartile range of -10.03% to 5.73%. This gives a nice visual representation of the data and gives further credence to the notion that the higher overall confidence in the self-assessment can be related to lower cost variation and better cost estimation predictor.

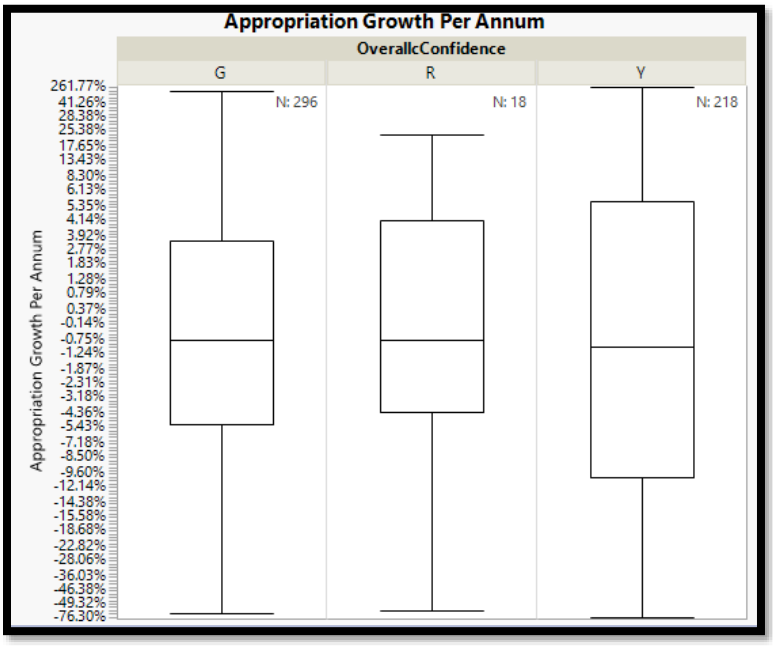


Figure 7: Box Plot Cost Growth Per Annum vs Overall Confidence

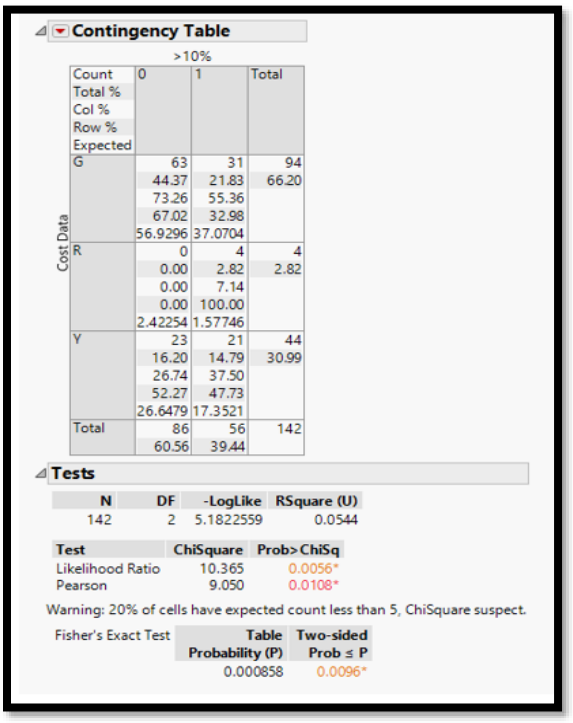
4.3 Which of the self-assessment confidence enablers are most predictive of future variance?

To answer which of the self-assessment measures are the most valuable as predictors of future, variance contingency tables were run on all the measures: requirements definition, engineering technical baseline, schedule baseline, cost data, crosschecks, risk assessment, and budget equals estimate. Each self-assessment measure was run by the dummy variable >10% cost variation. These measures where run against both production and development subsets out of the fourteen tests. Of the fourteen tests six produced significant results four in development and two in production discussed below. The complete results of all the contingency tables tests including the non-significant ones are included in appendix D.

The four measures of development with significance were: cost data, crosschecks, requirements, and schedule. The two measures of production with significant p-values they are: risk assessment and budget equals estimate.

Looking at the development phase testing first, the results for the "cost data" show the actual incidences for Green are greater than expected when cost variation is <10%. The dataset had 63 occurrences greater than the expected value of 56.9 occurrences. Yellow on the other hand had 31 occurrences compared to an expected value of 37.1 Green has more occurrences than expected when cost variance is lower than 10% and Yellow had more occurrence than expect when cost variance is higher than 10%. Therefore, we can determine the variables are dependent, that is cost data overall confidence is different when the cost variation is >10%. The p-value for the Pearson Test was statically significant at <.009 Table 7 below.

Table 7: Development Cost Data Measure vs >10% Cost Variance



Contingency Table

Cost Data	>10%	0	1	Total
	Count			
	Total %			
	Col %			
	Row %			
	Expected			
	G	63	31	94
		44.37	21.83	66.20
		73.26	55.36	
		67.02	32.98	
		56.9296	37.0704	
	R	0	4	4
		0.00	2.82	2.82
		0.00	7.14	
		0.00	100.00	
		2.42254	1.57746	
	Y	23	21	44
		16.20	14.79	30.99
		26.74	37.50	
		52.27	47.73	
		26.6479	17.3521	
	Total	86	56	142
		60.56	39.44	

Tests

N	DF	-LogLike	RSquare (U)
142	2	5.1822559	0.0544

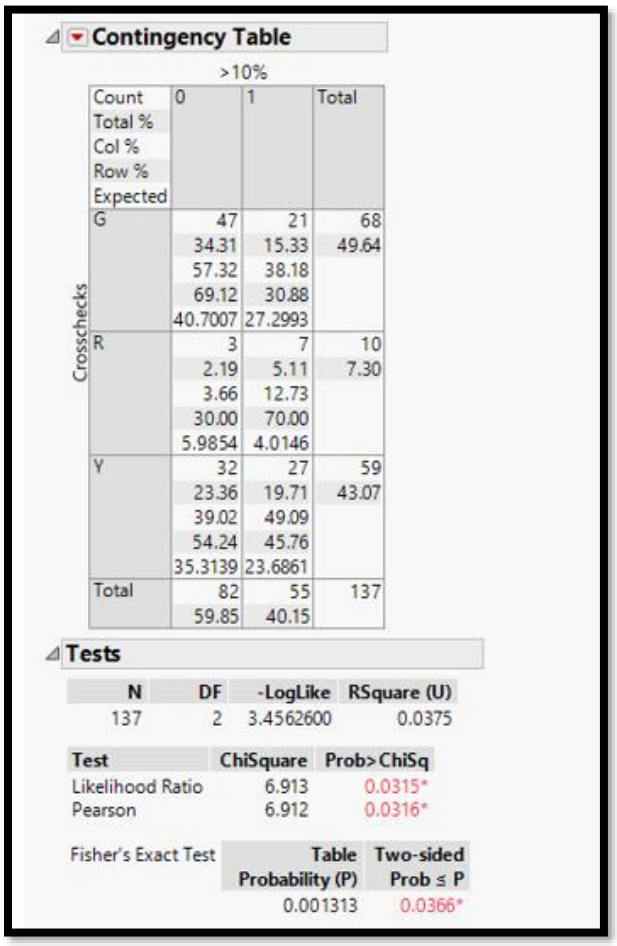
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.365	0.0056*
Pearson	9.050	0.0108*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000858	0.0096*

The results for "crosschecks" show the actual values for Green are greater than the expected ones when cost variation is <10%. The dataset had 47 occurrences -- greater than the expected value of 40.7 occurrences. Yellow on the other hand had 21 occurrences compared to an expected value of 27.3 Green has more occurrences than expected when cost variance is lower than 10% and Yellow had more occurrence than expect when cost variance is higher than 10%. The p-value for the Pearson Test was statically significant at <.036 Table 8 below.

Table 8: Development Crosscheck Measure vs >10% Cost Variance



Contingency Table

Crosschecks	>10% 0	>10% 1	Total
G	47	21	68
	34.31	15.33	49.64
	57.32	38.18	
	69.12	30.88	
	40.7007	27.2993	
R	3	7	10
	2.19	5.11	7.30
	3.66	12.73	
	30.00	70.00	
	5.9854	4.0146	
Y	32	27	59
	23.36	19.71	43.07
	39.02	49.09	
	54.24	45.76	
	35.3139	23.6861	
Total	82	55	137
	59.85	40.15	

Tests

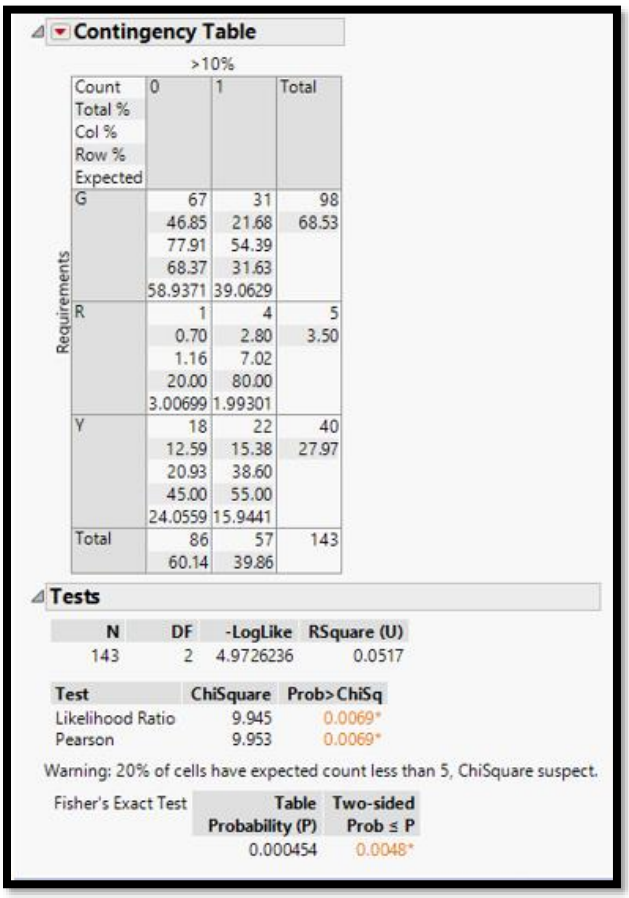
N	DF	-LogLike	RSquare (U)
137	2	3.4562600	0.0375

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.913	0.0315*
Pearson	6.912	0.0316*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.001313	0.0366*

"Requirements development" show the actual values for Green are greater than the expected ones when cost variation is <10%. The dataset had 67 occurrences greater than the expected value of 58.9 occurrences. Yellow had 31 occurrences compared to an expected value of 39.1 Green has more occurrences than expected when cost variance is lower than 10% and Yellow had more occurrence than expect when cost variance is higher than 10%. The p-value for the Pearson Test was statically significant at <.005 Table 9 below.

Table 9: Development Requirements Measure vs >10% Cost Variance



Contingency Table

Requirements	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		67	31	98
		46.85	21.68	68.53
		77.91	54.39	
		68.37	31.63	
		58.9371	39.0629	
R		1	4	5
		0.70	2.80	3.50
		1.16	7.02	
		20.00	80.00	
		3.00699	1.99301	
Y		18	22	40
		12.59	15.38	27.97
		20.93	38.60	
		45.00	55.00	
		24.0559	15.9441	
Total		86	57	143
		60.14	39.86	

Tests

N	DF	-LogLike	RSquare (U)
143	2	4.9726236	0.0517

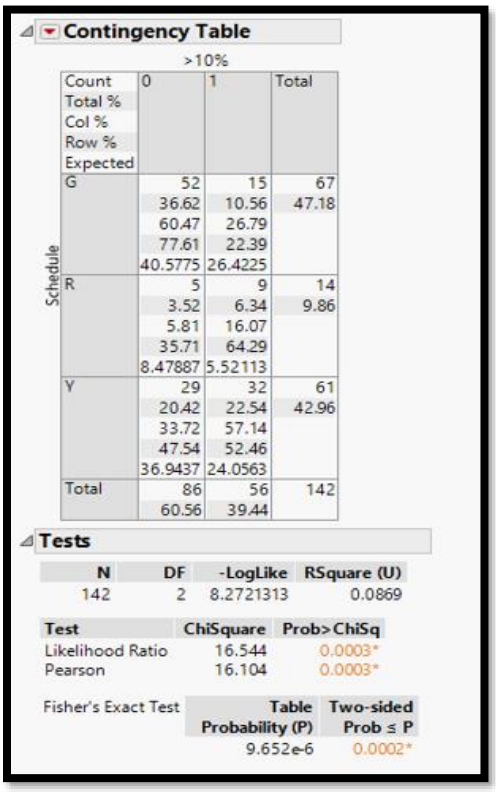
Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	9.945	0.0069*
Pearson	9.953	0.0069*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000454	0.0048*

Finally, "schedule" shows the actual values for Green are greater than the expected ones when cost variation is <10%. The dataset had 52 occurrences -- greater than the expected value of 40.58 occurrences. Yellow on the other hand had 15 occurrences compared to an expected value of 26.42 Green has more occurrences than expected when cost variance is lower than 10% and Yellow had more occurrence than expect when cost variance is higher than 10%. The p-value for the Pearson Test was statically significant at <.0002 Table 10 below. It is apparent that some categories have no statistical significance or effect on overall confidence, conclusions of them as a whole will be done in Chapter 5.

Table 10: Development Schedule Measure vs >10% Cost Variance



Contingency Table

Schedule	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		52	15	67
		36.62	10.56	47.18
		60.47	26.79	
		77.61	22.39	
		40.5775	26.4225	
R		5	9	14
		3.52	6.34	9.86
		5.81	16.07	
		35.71	64.29	
		8.47887	5.52113	
Y		29	32	61
		20.42	22.54	42.96
		33.72	57.14	
		47.54	52.46	
		36.9437	24.0563	
Total		86	56	142
		60.56	39.44	

Tests

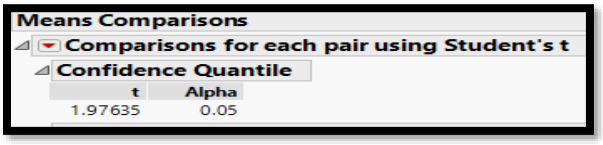
N	DF	-LogLike	RSquare (U)
142	2	8.2721313	0.0869

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	16.544	0.0003*
Pearson	16.104	0.0003*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	9.652e-6	0.0002*

Now we will discuss the results for the two procurement measures. The first measure for procurement is "risk assessment." The difference here from the previous tests is due to the Green and Yellow results being closer to what they were expected their means were not significantly different. A Student's t test confirmed this Table 11 below. It also shows that difference in Red's means causes the statistical significance in Schedule. The significant difference is only related to the incidences of red. Here the dataset had 1 red occurrence when expected when cost variance was lower than 10% less than the expected value of 4.16 occurrences and when cost variance was higher than 10% there where 6 occurrences when only 2.84 where expected. The Pearson Test was statically significant at <.0397 Table 12 below.

Table 11: Production Risk Assessment Measure vs >10% Cost Variance Student's Test

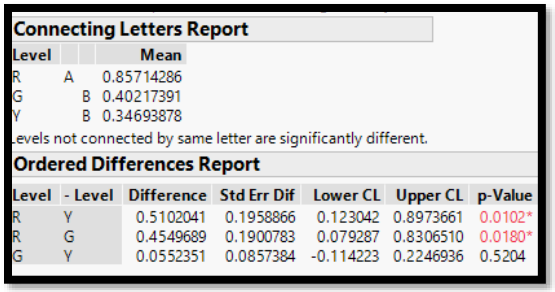


Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
1.97635	0.05



Connecting Letters Report

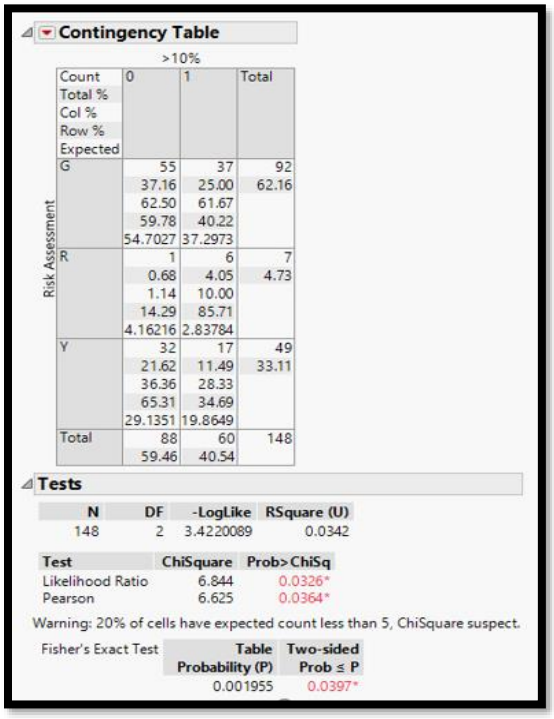
Level		Mean
R	A	0.85714286
G	B	0.40217391
Y	B	0.34693878

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
R	Y	0.5102041	0.1958866	0.123042	0.8973661	0.0102*
R	G	0.4549689	0.1900783	0.079287	0.8306510	0.0180*
G	Y	0.0552351	0.0857384	-0.114223	0.2246936	0.5204

Table 12: Production Risk Assessment Measure vs >10% Cost Variance



Contingency Table

Risk Assessment	>10%	0	1	Total
	Count			
	Total %			
	Col %			
	Row %			
	Expected			
G		55	37	92
		37.16	25.00	62.16
		62.50	61.67	
		59.78	40.22	
		54.7027	37.2973	
R		1	6	7
		0.68	4.05	4.73
		1.14	10.00	
		14.29	85.71	
		4.16216	2.83784	
Y		32	17	49
		21.62	11.49	33.11
		36.36	28.33	
		65.31	34.69	
		29.1351	19.8649	
Total		88	60	148
		59.46	40.54	

Tests

N	DF	-LogLike	RSquare (U)
148	2	3.4220089	0.0342

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.844	0.0326*
Pearson	6.625	0.0364*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.001955	0.0397*

Finally, the results for the "budget equals estimate" show the actual values for Yellow are greater than the expected ones when cost variation is <10%. Again the

statistically significant production results are not driven by solely by Green. The dataset had 46 occurrences greater than the expected value of 37.03 occurrences. Red, on the other hand, had 14 occurrences compared to an expected value of 19.7 Yellow had more occurrences than expected when cost variance is lower than 10% and Red had more occurrence than expect when cost variance is higher than 10%. A Student's t test confirmed this Table 14 shows that Yellow is statistically different from Green and Red The p-value for the Pearson Test was statically significant at <.005 Table 12 below.

Table 13: Production Budget Equals Estimate Measure vs >10% Cost Variance

Contingency Table

Budget Equals Estimate	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		29	25	54
		19.46	16.78	36.24
		32.58	41.67	
		53.70	46.30	
		32.255	21.745	
R		14	19	33
		9.40	12.75	22.15
		15.73	31.67	
		42.42	57.58	
		19.7114	13.2886	
Y		46	16	62
		30.87	10.74	41.61
		51.69	26.67	
		74.19	25.81	
		37.0336	24.9664	
Total		89	60	149
		59.73	40.27	

Tests

N	DF	-LogLike	RSquare (U)
149	2	5.2600107	0.0524

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.520	0.0052*
Pearson	10.317	0.0058*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000136	0.0050*

Table 14: Production Budget Equals Estimate Measure vs >10% Cost Variance

Student's Test

Means Comparisons

Comparisons for each pair using Student's t

Confidence Quantile

t	Alpha
1.97635	0.05

Connecting Letters Report

Level		Mean
R	A	0.57575758
G	A	0.46296296
Y	B	0.25806452

Levels not connected by same letter are significantly different.

Ordered Differences Report

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
R	Y	0.3176931	0.1029981	0.114133	0.5212530	0.0024*
G	Y	0.2048984	0.0889727	0.029058	0.3807392	0.0227*
R	G	0.1127946	0.1056151	-0.095937	0.3215266	0.2873

Analysis of these results shows the relationship between the each of the six measures and cost variation >10% were statically significant. Four all of the development measures had a greater occurrence of Yellow or Red confidence level than would be expected. This shows there is a higher chance of variation >10% for each of these measures with the lower confidence levels Red and Yellow self-assessments in development.

Breaking down the four development measures: cost data, crosschecks, requirements, and schedule. "Requirements" and "schedule" are significant predictors. From the Program office Pedigree instruction (Appendix A): Green rated "requirements" measure means the requirements is complete, stable, and well delineated, compared to Yellow rated requirements measure that are still in flux and may need assumptions to provide enough information for a complete estimate. Schedule would seem to be a significant

predictor of high cost variation was well. The difference between Green and Yellow rated schedules is defined by well-documented integrated program schedule versus a documented schedule which models some program risks. The two measures of production with significant p-values are: risk assessment, budget equals estimate. A possible explanation for a lower risk assessment is due to limited cost, schedule, and technical risk assessment modeled into cost estimate. The existence of these conditions would drive up cost variation. As for Budget equals estimate lower confidence level is defined by some out-year budget disconnects with mitigation plans. Again this can increase the variation in cost estimates.

The data supports using the measures with significant results from the program office self-assessment as predictors. The six measures that were statistically significant all show there is a difference when measures are assessed as Green, Yellow, or Red. Development in particular showed the measures: cost data, crosschecks, requirements, and schedule are significantly more likely to be Yellow or Red when the cost variance > 10% than if it is less than 10%.

4.4 When does a program office self-assessment peak in confidence during a program? When does confidence stabilize in a program?

To answer the research question, four contingency table analyses were used to determine peaks periods. The data set was broken down into development and production. Contingency table compared the overall confidence compared to percent work complete. The overall confidence level was analyzed by the second percent work

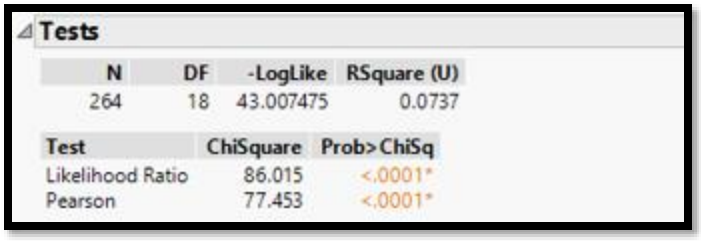
complete dummy variable. The dummy variable as discussed in Chapter III breaks the work complete into 10% intervals from 0-10% to 90-100%.

The data was different for the two subsets and accentuates the differences of the two stages. First, development Table 13 below shows the overall confidence didn't peat until the 90% complete stage. Here 90.48% of the overall confidence occurrences were Green. Interestingly from 40%-80% work complete the overall confidence Green occurred at around 60% with no real change or improvement in the development subset. The Pearson test p-value of <.0001 shows that programs overall confidence is significantly affected by the percent work completed.

Table 15: Development Overall Confidence vs Percent Work Complete-10% Intervals

Contingency Table

OverallConfidence	Count / Total % / Col % / Row % / Expected	DV_WK2 1	2	3	4	5	6	7	8	9	10	Total
G		7	9	8	11	16	16	11	6	19	40	143
		2.65	3.41	3.03	4.17	6.06	6.06	4.17	2.27	7.20	15.15	54.17
		14.00	45.00	38.10	57.89	57.14	64.00	45.83	60.00	90.48	86.96	
		4.90	6.29	5.59	7.69	11.19	11.19	7.69	4.20	13.29	27.97	
		27.0833	10.8333	11.375	10.2917	15.1667	13.5417	13	5.41667	11.375	24.9167	
R		6	0	1	1	0	1	0	0	0	0	9
		2.27	0.00	0.38	0.38	0.00	0.38	0.00	0.00	0.00	0.00	3.41
		12.00	0.00	4.76	5.26	0.00	4.00	0.00	0.00	0.00	0.00	
		66.67	0.00	11.11	11.11	0.00	11.11	0.00	0.00	0.00	0.00	
		1.70455	0.68182	0.71591	0.64773	0.95455	0.85227	0.81818	0.34091	0.71591	1.56818	
Y		37	11	12	7	12	8	13	4	2	6	112
		14.02	4.17	4.55	2.65	4.55	3.03	4.92	1.52	0.76	2.27	42.42
		74.00	55.00	57.14	36.84	42.86	32.00	54.17	40.00	9.52	13.04	
		33.04	9.82	10.71	6.25	10.71	7.14	11.61	3.57	1.79	5.36	
		21.2121	8.48485	8.90909	8.06061	11.8788	10.6061	10.1818	4.24242	8.90909	19.5152	
Total		50	20	21	19	28	25	24	10	21	46	264
		18.94	7.58	7.95	7.20	10.61	9.47	9.09	3.79	7.95	17.42	



Tests

N	DF	-LogLike	RSquare (U)
264	18	43.007475	0.0737

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	86.015	<.0001*
Pearson	77.453	<.0001*

Table 14 below shows the production overall confidence subset. It actually peaks at 60% work complete with Green overall confidence occurring at 92.31%. Procurement shows program offices are more confident than they are at similar stages in the development work complete. Procurement Green overall confidence is at 68.41% when the program has 20% of the work completed. The Pearson test p-value of <.0001 shows that programs overall confidence is significantly affected by the percent work completed.

Table 16: Production Overall Confidence vs Percent Work Complete-10% Intervals

Contingency Table

OverallConfidence	Count / Total % / Col % / Row % / Expected	DV_WK2 1	2	3	4	5	6	7	8	9	10	Total
G		54	13	12	11	10	12	13	8	7	13	153
		20.15	4.85	4.48	4.10	3.73	4.48	4.85	2.99	2.61	4.85	57.09
		39.71	68.42	66.67	68.75	71.43	92.31	81.25	61.54	87.50	86.67	
		35.29	8.50	7.84	7.19	6.54	7.84	8.50	5.23	4.58	8.50	
		77.6418	10.847	10.2761	9.13433	7.99254	7.42164	9.13433	7.42164	4.56716	8.56343	
R		7	1	1	0	0	0	0	0	0	0	9
		2.61	0.37	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.36
		5.15	5.26	5.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
		77.78	11.11	11.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
		4.56716	0.63806	0.60448	0.53731	0.47015	0.43657	0.53731	0.43657	0.26866	0.50373	
Y		75	5	5	5	4	1	3	5	1	2	106
		27.99	1.87	1.87	1.87	1.49	0.37	1.12	1.87	0.37	0.75	39.55
		55.15	26.32	27.78	31.25	28.57	7.69	18.75	38.46	12.50	13.33	
		70.75	4.72	4.72	4.72	3.77	0.94	2.83	4.72	0.94	1.89	
		53.791	7.51493	7.1194	6.32836	5.53731	5.14179	6.32836	5.14179	3.16418	5.93284	
Total		136	19	18	16	14	13	16	13	8	15	268
		50.75	7.09	6.72	5.97	5.22	4.85	5.97	4.85	2.99	5.60	

Tests

N	DF	-LogLike	RSquare (U)
268	18	23.505808	0.0497

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	47.012	0.0002*
Pearson	41.569	0.0013*

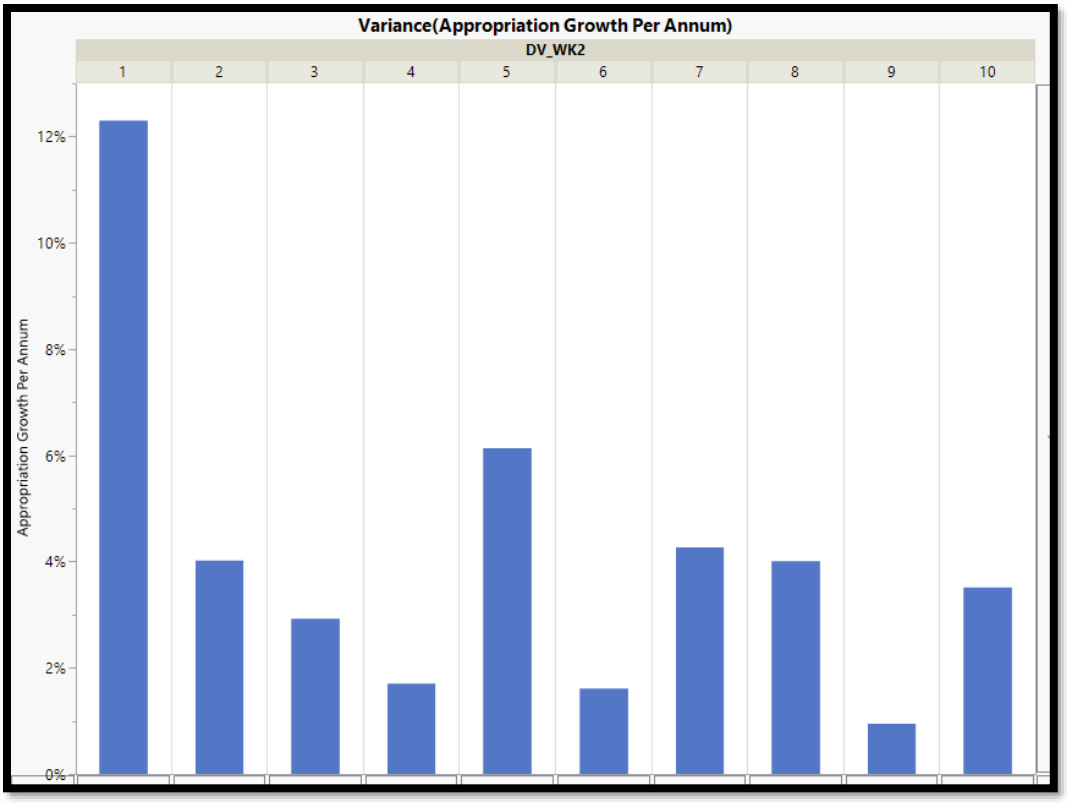


Figure 8: Cost Growth Per Annum vs Percent Work Complete-10% Intervals

Figure 8 supports the result of Tables 13 and 14. The first 10% work complete has 12% variance. By 20% work complete the variance drops to 4% and while it has a few spikes the greatest variance in in the first 10% of work complete. It supports self-assessments improving as the greater amount of work is completed and as programs progress the variation in cost growth drops.

Analysis of the results show program office self-assessments peak at different time for development and procurement. It 20% work complete procurement had a Green overall confidence occurrence of 68.41% for development this occurrence did not happen until the work complete was a 80%. So program offices are far more confident in the procurement stage than the development one. Within each, program offices where most

confident in development at 80% work complete or once this stage was practically finished. Procurement was not only more confident early but it peaked are 60% work complete. This is significant because there is still a fair amount of work to be completed after this point, but they are much more confident much earlier for procurement.

4.5 To what extent can we attribute changes in program estimates controllable factors?

The answer the fifth research question contingency tables analysis was used. The contingency table analysis was done with the dummy variable >10% cost variation run against the variable "reason for change" in program office estimates as discussed in the Chapter III methodology. The overall confidence levels of Green, Red, and Yellow were analyzed against the occurrences of the reason for change: quantity, schedule, engineering, estimating, economic, scope, and support.

Table 17: Cost Growth >10% vs Reason for Change

Contingency Table

DV >10%	ReasonForChange						
Count Total % Col % Row % Expected	Economic	Engineering Change Order	Estimating	Quantity	Requirements/ scope change	Schedule	Total
0	3 2.40 60.00 4.92 2.44	1 0.80 50.00 1.64 0.976	43 34.40 61.43 70.49 34.16	1 0.80 11.11 1.64 4.392	11 8.80 35.48 18.03 15.128	2 1.60 25.00 3.28 3.904	61 48.80
1	2 1.60 40.00 3.13 2.56	1 0.80 50.00 1.56 1.024	27 21.60 38.57 42.19 35.84	8 6.40 88.89 12.50 4.608	20 16.00 64.52 31.25 15.872	6 4.80 75.00 9.38 4.096	64 51.20
Total	5 4.00	2 1.60	70 56.00	9 7.20	31 24.80	8 6.40	125

Tests

N	DF	-LogLike	RSquare (U)
125	5	7.3802949	0.0480

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	14.761	0.0114*
Pearson	13.850	0.0166*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	2.658e-6	0.0080*

The results show the actual values for "estimating" as a reason for change are less than expected when cost variation is >10%. Table 15 shows the dataset had only had 27 occurrences in the first quartile but an expected value of 35.84 occurrences. The occurrence of quantity, schedule, and scope all more than expected when cost variation is >10%. The p-value for the Fisher's Exact Test was statically significant at <.008

These results support the idea that SPOs or the inputs they control have less impact in larger variances POEs in cost estimation. From categories of quantity, schedule, engineering, estimating, economic, other, and support they only have total control of the estimating and even then this is dependent on the information provided by others. The data showed that if there is a variation in a cost estimate of >10% a significantly less portion of the changes were due to estimating than expected. In total estimating accounts for 56% of the POE changes, however this drops to 42% when there is a greater than >10% change. So while estimating errors may have the most occurrences as to the reason of a POE change, they are typically smaller and of less impact than the less frequent but greater cost growth impacting changes such as schedule or scope.

As mentioned in the lit review Bolten categorized the reasons for cost growth similar the program office reason for change. These variance categories along with how they classified the variance data showed the "realized" cost growth and decisions by the government dominate the overall growth in both development and procurement. (Bolten, 2008). Showing that "realized" cost growth is often out of the program offices hand.

The reoccurring theme of the Program office and the cost estimator being easy targets to blame when growth cost occurs due to the responsibility of accurate cost estimating is still valid. (DeNeve et al. 2015) However, while SPOs do have the greatest percentage of changes the data has shown the cost variation of the reasons within their control is significantly less than the reasons outside of their control. Can they continue to strive for improvement and better estimates? Yes, but also there is only so much that they can do given all of the other constraints as well. The issues are larger than the program office and are only part of the solution to a greater problem.

4.6 Have cost estimates improved over time within the SPO? And has the technique of self-assessment proved more valuable over time?

The answer the sixth research question first descriptive statistics were used. Cost estimates where graphed against the program office's the mean Appropriation Growth per Annum calculation. The descriptive statistics of the dataset where analyzed for: mean, variance, and quartiles. The POE dataset had a mean of -1.3% with a variance of 6.4% The negative mean showed the dataset had slightly negative average cost growth for its entirety. The dataset had tight upper and lower quartiles of 4.09% and -7.86. These results tell us there most of the data is very close to the mean, more than half of the estimates where within +/-5% of the mean.

Figure 9 shows a slight improvement of the total cost variation over time. Cost estimates are closer to the goal 0.0% estimation error absolutely. However, there were no

statistically significant results when the dataset was run for the mean Appropriation Growth per Annum.

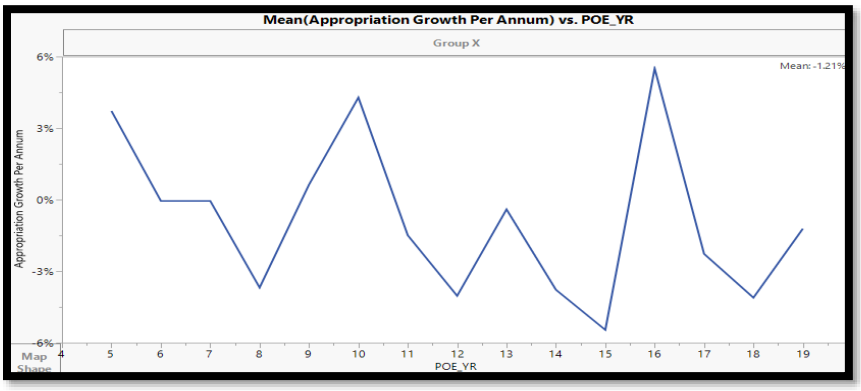


Figure 9: Total Cost Growth Per Annum vs POE Year

When divided out between development and procurement there are some differences. Figure 10 shows a slight regression of the total cost variation over time for development as the values spike and vary of time. This is somewhat understandable as the programs are far less predictable in the development stage. Next we will compare it to procurement.

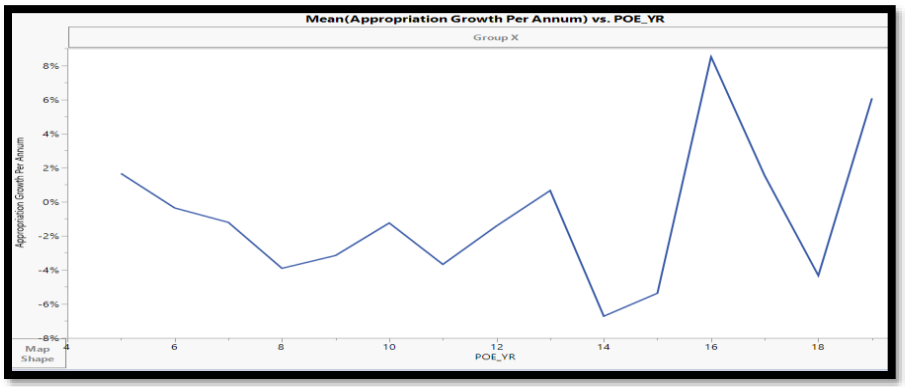


Figure 10: Development Cost Growth Per Annum vs POE Year

Figure 11 below shows the total cost variation in procurement. It has shown marked improvement overtime. Recent trends are to the point of potential under execution, but the program offices obviously have a stronger grasp of the programs at the procurement stage as well as they are less likely to be impacted by factors outside of their control.

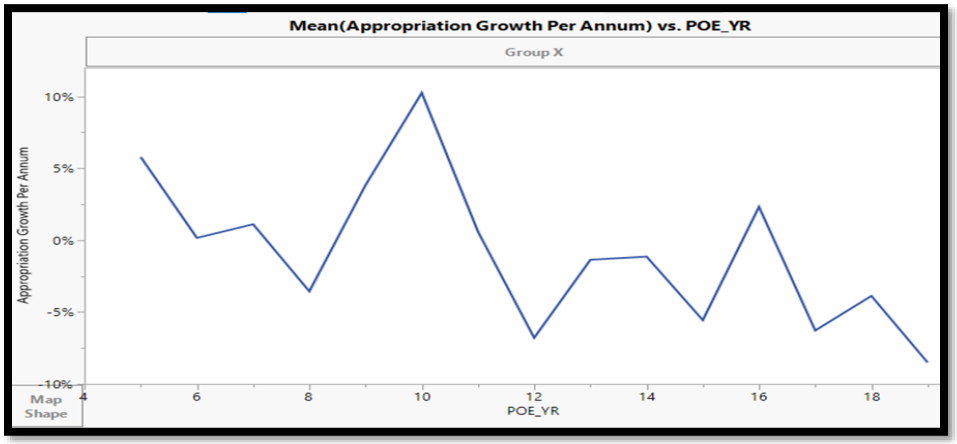


Figure 11: Production Cost Growth Per Annum vs POE Year

The results and other literature show cost estimates may have slight improvement over time within the SPO but nothing that is statistically significant. The data cannot confirm that estimates at the SPO have improved over time and the literature agrees with this as well. Younossi pointed out perhaps the most important finding of their analysis was cost growth in the past three decades has remained high, with no significant improvement. (Younossi, 2007) Even after reform WSARA stated new defense weapon systems programs have done a better job staying within budget estimates but proceed without the key knowledge essential to good acquisition outcomes (GAO, 2018)

Has the technique of self-assessment proved more valuable over time? There is value in self-assessment, as with any other factor being applied to cost growth. Self-assessment causes an organization to think and think critically. The Program offices can see how they truly are performing and give them a chance to evaluate their program critically. The literature has shown self-assessment tools can aid an organization provided they are proper procedures are followed on how to complete the assessment. Self-assessment can be defined as a 'learning process' organizations need more information before making a decision along with technical support and instruments to guide them.

Tying it to the SPOs specifically greater familiarity with the assessment criteria will provide value two-fold. First, it will allow the organization better to understand all the processes being utilized. And secondly it will allow them to see where the specific program stands and allow the organization to review programs to categorize for better priority or focus.

Summary

Chapter IV analyzed the results for the investigative questions with descriptive statistics and contingency table analysis. Additionally, the results and any significant findings were used to answer the investigative questions. Differences between green and yellow overall confidence assessments where found, however there was little explanatory power to the tests. Chapter V draws conclusions about Self-assessment and future research recommendations.

V. Conclusions and Recommendations

Chapter Overview

This chapter concludes form results the value and usefulness of self-assessment by discussing the limitations of the research, recommendation for future research, and final thoughts on self-assessment

Limitations

While the findings of this study had some significant results, limitations exist in this research. Data from is only from AFLCMC and their POEs deeper analysis into other agency's SPOs would be needed to better understand the value DoD wide. More data must be made available across all agencies to have better data sets and improve analysis.

Additionally, a Green, Yellow, Red tool has its limitations. By only having three categories to choose from there is little room for differentiation. While conducting a data collection, assessors wanted to add a level of Green-Yellow or Yellow-Red stratification to the self-assessment, but the Program Office Self-Assessment instructions only identify the three level of color coding. Furthermore, continued learning on POE submission requirements should ensure all submissions have a uniform self-assessment template to eliminate any variation.

Recommendations for Future Research

Further research could tie into specific cost growth methodologies within the POE database. Utilizing the self-assessment overall confidence to see if program offices are

more confident in certain methodologies. Additionally, specific program types or offices could be studied to see if there is any correlation between them and self-assessment or overall program confidence. Further study could be done with the self-assessments database and other program offices as there are limits to what we can learn within one SPO.

AFLCMC database has significant untapped resources with over seventeen years of historical program office data on PDF files. The POE files contain specific research materiel for example Technology Readiness Levels (TRL) within the POE briefing could be researched and coded for further analysis.

Final Thoughts

This thesis explored the nature of self-assessment. The literature has identified self-assessments as staple of many organizations has value. Self-assessments have the potential to be a useful tool in an organization. Self-assessment not only teach the members of an organization to think critically, it guides them while making decisions and generating solutions. When conducting an assessment, the assessor should always be aware of over confidence on the assessment.

The self-assessment may have overall value the specific measures can have individual values as well. There were four measures of development with significant results: cost data, crosschecks, requirements, and schedule. Closely tracking these measures, specifically requirements and schedule would be of value as they are indicators of greater cost variation. At the production stage risk assessment and budget equals estimate were significant. The fact that the measures are different signals a change in the self-

assessment from the development to the production stage. SPOs should now focus on budget estimation and mitigating risk. Closer inspection of these specific program measures could lead to improved overall confidence in the programs and as the data supports reduce the likelihood of cost variation.

Tying this back to program office estimates the self-assessment has a role in monitoring and tracking programs. There were significant differences between the variance of Green and Yellow Overall Confidence levels. So the value in knowing that the self-assessment can give a quick idea on how the program will perform. This stop light assessment presented is presented on the program office estimates. At this time the current state of the self-assessment tools seems the most warranted for the SPOs.

Appendix

Appendix A

Self-Assessment Program Confidence Criteria MS-A-Pre MS-B

PROGRAM "Confidence" PEDIGREE *"PROGRAM REALISM is needed to achieve COST REALISM"*			
Program Name:		**Program Phase: MS-A - Pre MS-B**	**Date:**
Confidence Enablers	***"HIGH" Confidence***	***"MEDIUM" Confidence***	***"LOW" Confidence***
Requirements Definition	- Complete, stable, user approved/coordinated set of high level functional and performance requirements that define system attributes - Few assumptions required - Documented program plan/acquisition strategy with sufficient detail for a comprehensive cost estimate	- Generally understood, some functional and performance requirements still in flux but not major cost drivers - Requirements change process with senior level gatekeeper in place - Assumptions required to provide enough information to complete estimate are provided by appropriate functional lead	- Requirements in deliberation - Multiple major assumptions required to complete estimate
Engineering Technical Baseline	The descriptions in the baseline support defining major hardware, software and integration elements and provide reasonable insights into greatest risk areas and cost drivers at the subsytem level. The system as a whole is well characterized; reference design is documented. Ground rules and assumptions are well documented. All DoD 5000.02 documents are leveraged.	The system is described is sufficient detail to support identification of high level analogies and application of parametric modeling. Many but not all DoD 5000.02 documents are leveraged. A high-level reference design with some assumptions forms the technical basis of the cost estimate.	The descriptions in the baseline are vague and incomplete, even at the system level. DoD 5000.02 documents are not leveraged. Significant uncertainty about system architecture and technical approach.
Schedule Baseline	- Comprehensive, detailed well documented integrated program schedule with durations, logic, and "what if" (impact on task duration) analysis completed for potential risks. Based on schedule analysis for the tasks required, that has been vetted with cost, design, engineering and test organizations. Coordinated with acquisition strategy	- Comprehensive but higher level documented schedule which models some program risks. Based on schedule analysis for individual tasks required, but lacks full integration. Coordinated with acquisition strategy.	- Top level notional schedule lacking definition to model program risks. Downward directed,based on user need date, NOT duration and complexity of tasks required. Lacks coordination with acquisition strategy.
Cost Data & Methodology	- CERs or analogies to similar programs used to estimate a at a level of detail consistent with tech baseline and requirements maturity. Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in AF Cost Analysis Handbook.	- CERs, commercial models, or analogies to less similar but like function programs used to estimate a at a level of detail consistent with tech baseline and requirements maturity. Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in AF Cost Analysis Handbook.	- Commercial parametric models or manpower build-up (lacking historical substantiation) used as primary methodologies. Limited use of analogous data.
Crosscheck(s)	- Estimate cost drivers have been crosschecked with historical/actual data on similar programs and are supportive of the program estimate results	- Few crosschecks available – generally support the estimate	- No appropriate cross-checks used or results do not support estimate
Risk Assessment (Cost/Schedule/ Tech)	- Comprehensive cost, schedule, and technical risk assessment modeled into cost estimate in compliance with JA Cost, Schedule, Risk, & Uncertainty Handbook, and cost estimating best practices. Cost estimate integrates identified program risks and the cost/phasing impact of these risks can be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle. Ranges of highest risk/uncertainty areas and cost drivers based on well documented technical baseline and vetted with appropriate engineering or functional OPRs.	- Limited cost, schedule, and technical risk assessment modeled into cost estimate but in compliance with JA Cost, Schedule, Risk, & Uncertainty Handbook. Cost estimate integrates identified program risks but the cost/phasing impact of these risks CANNOT be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle.	- Little or no cost, schedule, and technical risk assessment modeled into cost estimate or risk modeled based on generic assumptions by the cost estimator. Narrow CDF for a program at this phase in the lifecycle.
Budget *equals* Estimate	- Virtually no disconnects between program estimate and budget	- No near-term disconnects between estimate and budget, Some out-year disconnects with mitigation plans	- Major disconnects between program Estimate and budget
Overall Assessment	Provide an overall assessment of the program Estimate Confidene. No program with a "red" in any category can be assessed higher than a yellow, "yellow" and "red" results must be documented in findings.		

Appendix B

Self-Assessment Program Confidence Criteria MS-B-Pre MS-C

PROGRAM "Confidence" PEDIGREE *"PROGRAM REALISM is needed to achieve COST REALISM"*			
Program Name:		**Program Phase: Approaching MS-B - MS-C**	**Date:**
Confidence Enablers	***"HIGH" Confidence***	***"MEDIUM" Confidence***	***"LOW" Confidence***
Requirements Definition	- Complete, stable, user approved/coordinated set of requirements - (well delineated and cross-referenced CDD and SRD; APB in place). Requirements management strictly controlled; changes reflected in cost updates) Few assumptions required - Documented program plan/acquisition strategy with sufficient detail for a comprehensive cost estimate	- Generally understood, some areas still in flux but not major cost drivers - Requirements change process with senior level gatekeeper in place - Assumptions required to provide enough information to complete estimate but provided by appropriate functional lead; Able to assess cost and schedule risk associated with requirements uncertainty or instability.	- Requirements in deliberation - Multiple major assumptions required to complete estimate
Engineering Technical Baseline	The descriptions are comprehensive and costable at the subsystem level. Integration and test at subsystem and system levels defined. Program strategies (technology development, acquisition/contracting, etc) are summarized and assessed to enable evaluation of cost and risk impact. Non-mission equipment cost elements are described for all life cycle phases. The baseline reflects all recent changes to program strategies. All DoD 5000.02 documents are fully leveraged.	Subsystems and non-mission equipment elements are described in enough detail to support identification and cost analysis of key risks and all cost drivers. All DoD 5000.02 documents are fully leveraged.	At the subsystem level, descriptions are incomplete and vague. Many 5000.02 documents ignored or underutilized. Or, the program merely references contractor's knowledge base without maintaining an independent, objective, and up-to-date description of the product baseline and all cost elements.
Schedule Baseline	- Comprehensive, detailed well documented integrated program schedule with durations, logic, and "what if" (impact on task duration) analysis completed for potential risks. Based on schedule analysis for the tasks required. Coordinated with contractor IMS.	- Comprehensive, high level, documented schedules which may, or may not, be fully integrated . Coordinated with contractor IMS.	- Top level notional schedule lacking definition to model program risks. Based on user need date, NOT duration of tasks required.
Cost Data & Methodology	- Actuals on same program or very analogous program and/or parametric data at a comprehensive level of detail for virtually every Level 3 WBS element. Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in AF Cost Analysis Handbook.	- Analogous and/or parametric data on somewhat relevant programs for most Level 3 WBS elements Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in AF Cost Analysis Handbook.	- Commercial parametric models or manpower build-up (lacking historical substantiation) used as primary methodologies. Limited use of analogous data.
Crosscheck(s)	- Estimate cost drivers have been crosschecked with historical/actual data on similar programs and are supportive of the program estimate results	- Few crosschecks available – generally support the estimate	- No appropriate cross-checks used or results do not support estimate
Risk Assessment (Cost/Schedule/ Tech)	- Comprehensive cost, schedule, and technical risk assessment modeled into cost estimate in compliance with JA Cost, Schedule, Risk, & Uncertainty Handbook, and cost estimating best practices. Cost estimate integrates identified program risks and the cost/phasing impact of these risks can be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle. Ranges of highest risk/uncertainty areas and cost drivers based on well documented technical baseline and vetted with appropriate engineering or functional OPRs.	- Limited cost, schedule, and technical risk assessment modeled into cost estimate but in compliance with JA Cost, Schedule, Analysis, Risk, & Uncertainty Handbook. Cost estimate integrates identified program risks but the cost/phasing impact of these risks CANNOT be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle.	- Little or no cost, schedule, and technical risk assessment modeled into cost estimate or risk modeled based on generic assumptions by the cost estimator. Narrow CDF for a program at this phase in the lifecycle.
Budget *equals* Estimate	- Virtually no disconnects between program estimate and budget	- No near-term disconnects between estimate and budget, Some out-year disconnects with mitigation plans	- Major disconnects between program Estimate and budget
Overall Assessment	Provide an overall assessment of the program Estimate Confidene. No program with a "red" in any category can be assessed higher than a yellow, "yellow" and "red" results must be documented in findings.		

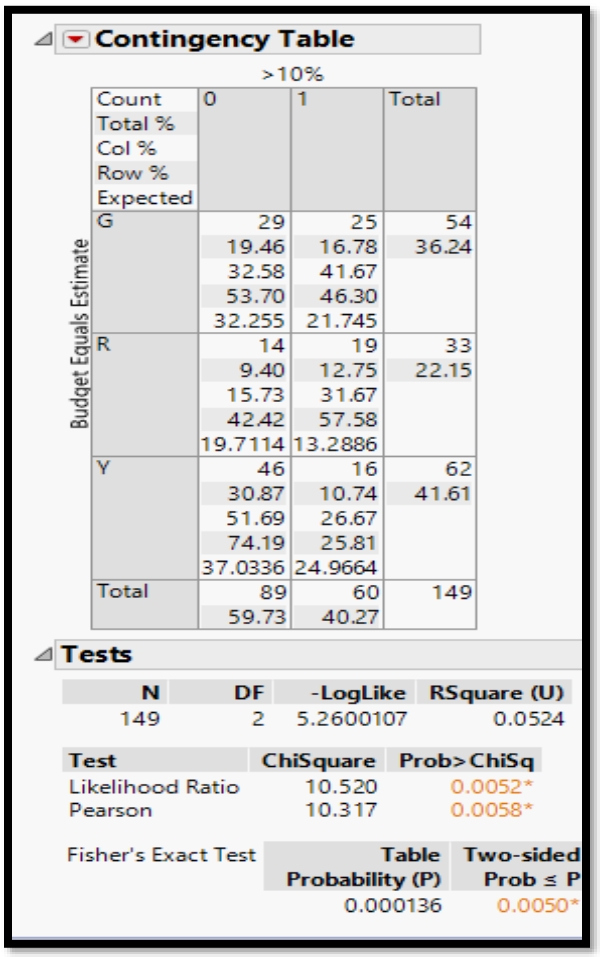
Appendix C

Self-Assessment Program Confidence Criteria Post MS-C

PROGRAM "Confidence" PEDIGREE *"PROGRAM REALISM is needed to achieve COST REALISM"*			
Program Name:		**Program Phase: Post MS-C**	**Date:**
Confidence Enablers	***"HIGH" Confidence***	***"MEDIUM" Confidence***	***"LOW" Confidence***
Requirements Definition	- Complete, stable, user approved/coordinated set functional/performance requirements verified by DT&E and OT&E (CPD and product baseline stable). Required delivery/deployment quantities and schedules clear.- Unmet/deferred requirements and strategy for future capability increments/blocks/releases clearly delineated. - Documented program plan/acquisition strategy with sufficient detail for a comprehensive cost estimate	- Well understood, including any requirements changes likely to impact production configuration and possibly require reintegration/test in development environment- Requirements change process with senior level gatekeeper in place - Assumptions required to provide enough information to complete estimate and assess risk, provided by appropriate functional lead	- Requirements in deliberation likely to cause undefined change and delay in production/deployment configuration - Multiple major assumptions required to complete estimate
Engineering Technical Baseline	Descriptions of mission equipment and collateral cost elements are comprehensiove and costable at the component level. The baseline reflects all recent changes to the sustainment approach, etc. All DoD 5000.02 documents fully leveraged.	Descriptions are comprehensive and costable, but only at the subsystem level. All DoD 5000.02 documents are fully leveraged. There is some user iteration, but limited.	The baseline lacks comprehensiveness and details approprate for this level of maturity. Or, program merely references contractor's knowledge base without maintaining an independent, objective, and up-to-date description of the product baseline and all cost elements.
Schedule Baseline	- Comprehensive, detailed well documented integrated program schedule with durations, logic, and "what if" (impact on task duration) analysis completed for potential risks. Based on schedule analysis for the tasks required. Coordinated with contractor IMS. Delivery schedule also coordinated with user install/deployment needs.	- Comprehensive, high level, documented schedules which may, or may not, be fully integrated . Coordinated with contractor IMS.	- Top level notional schedule lacking definition to model program risks. Based on user need date, NOT duration of tasks required.
Cost Data & Methodology	- Actuals on same program or very analogous program and/or parametric data at a comprehensive level of detail for virtually every Level 3 WBS element. Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in JA Cost, Schedule, Risk, & Uncertainty Handbook.	- Analogous and/or parametric data on somewhat relevant programs for most Level 3 WBS elements Estimators re-validated any re-used methodologies from other estimates for relevance. Estimating consistent with guidance in JA Cost, Schedule, Risk, & Uncertainty Handbook.	- Commercial parametric models or manpower build-up (lacking historical substantiation) used as primary methodologies. Limited use of analogous data.
Crosscheck(s)	- Estimate cost drivers have been crosschecked with historical/actual data on similar programs and are supportive of the program estimate results	- Few crosschecks available – generally support the estimate	- No appropriate cross-checks used or results do not support estimate
Risk Assessment (Cost/Schedule/ Tech)	- Comprehensive cost, schedule, and technical risk assessment modeled into cost estimate in compliance with JA Cost, Schedule, Risk & Uncertainty Handbook, and cost estimating best practices. Cost estimate integrates identified program risks and the cost/phasing impact of these risks can be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle. Ranges of highest risk/uncertainty areas and cost drivers based on well documented technical baseline and vetted with appropriate engineering or functional OPRs.	- Limited cost, schedule, and technical risk assessment modeled into cost estimate but in compliance with JA Cost, Schedule, Risk, & Uncertainty Handbook. Cost estimate integrates identified program risks but the cost/phasing impact of these risks CANNOT be discreetly identified; Sufficiently broad CDF for a program at this stage of it's lifecycle.	- Little or no cost, schedule, and technical risk assessment modeled into cost estimate or risk modeled based on generic assumptions by the cost estimator. Narrow CDF for a program at this phase in the lifecycle.
Budget *equals* Estimate	- Virtually no disconnects between program estimate and budget	- No near-term disconnects between estimate and budget, Some out-year disconnects with mitigation plans	- Major disconnects between program Estimate and budget
Overall Assessment	Provide an overall assessment of the program Estimate Confidene. No program with a "red" in any category can be assessed higher than a yellow, "yellow" and "red" results must be documented in findings.		

Appendix D

Table D1 Contingency Table Results Development Budget Equals Estimate Measure vs >10% Cost Variance



Contingency Table

Budget Equals Estimate	>10%	0	1	Total
	Count			
	Total %			
	Col %			
	Row %			
	Expected			
	G	29	25	54
		19.46	16.78	36.24
		32.58	41.67	
		53.70	46.30	
		32.255	21.745	
	R	14	19	33
		9.40	12.75	22.15
		15.73	31.67	
		42.42	57.58	
		19.7114	13.2886	
	Y	46	16	62
		30.87	10.74	41.61
		51.69	26.67	
		74.19	25.81	
		37.0336	24.9664	
	Total	89	60	149
		59.73	40.27	

Tests

N	DF	-LogLike	RSquare (U)
149	2	5.2600107	0.0524

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.520	0.0052*
Pearson	10.317	0.0058*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000136	0.0050*

Table D2 Contingency Table Results Development Cost Data Measure vs >10% Cost Variance

Contingency Table

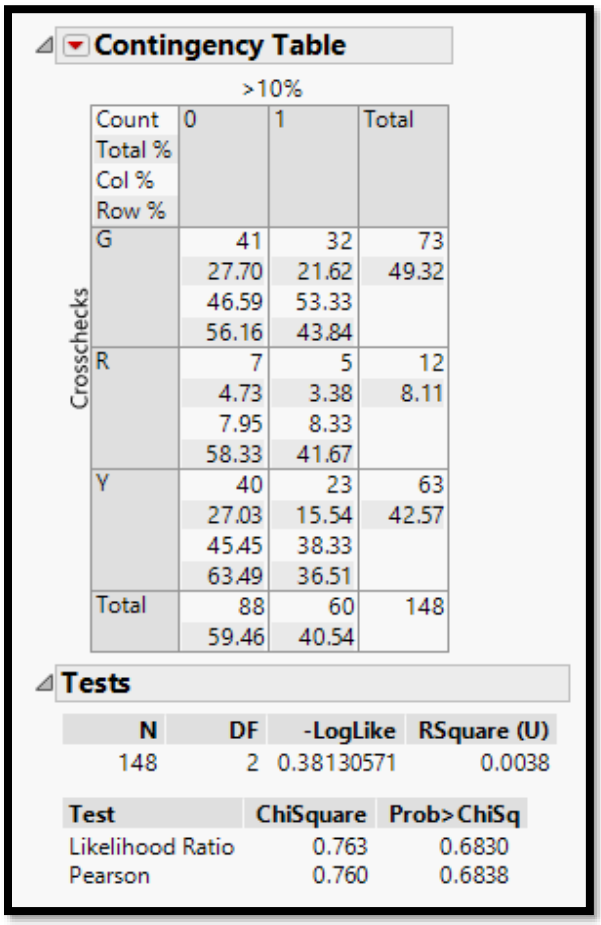
Cost Data	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		60	47	107
		39.47	30.92	70.39
		66.67	75.81	
		56.07	43.93	
R		3	1	4
		1.97	0.66	2.63
		3.33	1.61	
		75.00	25.00	
Y		27	14	41
		17.76	9.21	26.97
		30.00	22.58	
		65.85	34.15	
Total		90	62	152
		59.21	40.79	

Tests

N	DF	-LogLike	RSquare (U)
152	2	0.81816538	0.0080

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	1.636	0.4412
Pearson	1.598	0.4498

Table D3 Contingency Table Results Development Crosscheck Measure vs >10% Cost Variance



Contingency Table

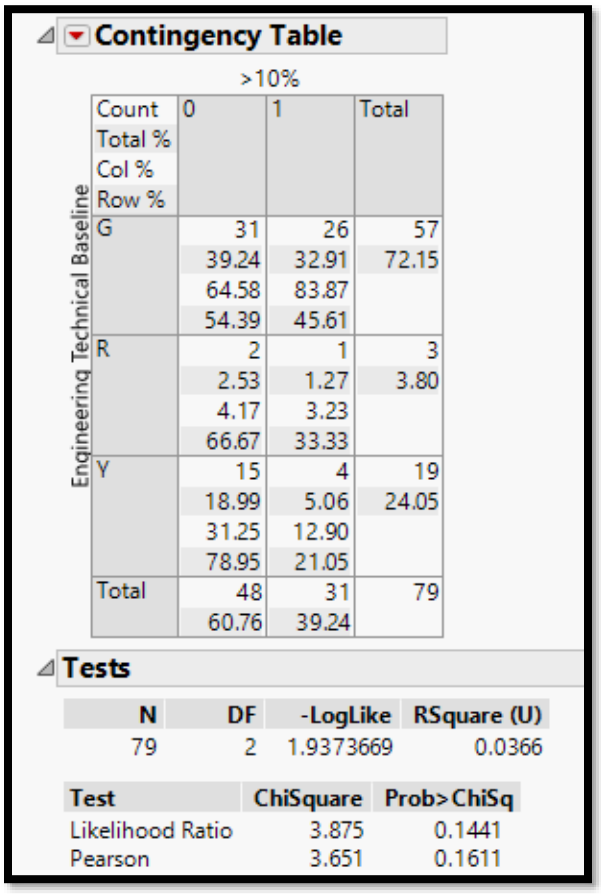
Crosschecks	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		41	32	73
		27.70	21.62	49.32
		46.59	53.33	
		56.16	43.84	
R		7	5	12
		4.73	3.38	8.11
		7.95	8.33	
		58.33	41.67	
Y		40	23	63
		27.03	15.54	42.57
		45.45	38.33	
		63.49	36.51	
Total		88	60	148
		59.46	40.54	

Tests

N	DF	-LogLike	RSquare (U)
148	2	0.38130571	0.0038

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	0.763	0.6830
Pearson	0.760	0.6838

Table D4 Contingency Table Results Development Engineering Technical Baseline Measure vs >10% Cost Variance



Contingency Table

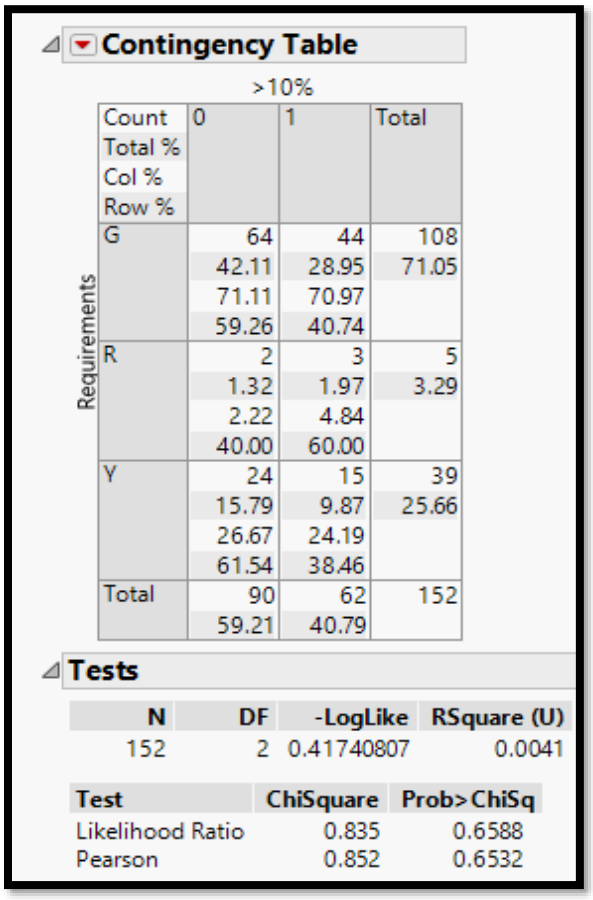
Engineering Technical Baseline	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		31	26	57
		39.24	32.91	72.15
		64.58	83.87	
		54.39	45.61	
R		2	1	3
		2.53	1.27	3.80
		4.17	3.23	
		66.67	33.33	
Y		15	4	19
		18.99	5.06	24.05
		31.25	12.90	
		78.95	21.05	
Total		48	31	79
		60.76	39.24	

Tests

N	DF	-LogLike	RSquare (U)
79	2	1.9373669	0.0366

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	3.875	0.1441
Pearson	3.651	0.1611

Table D5 Contingency Table Results Development Requirements Measure vs >10% Cost Variance



Contingency Table

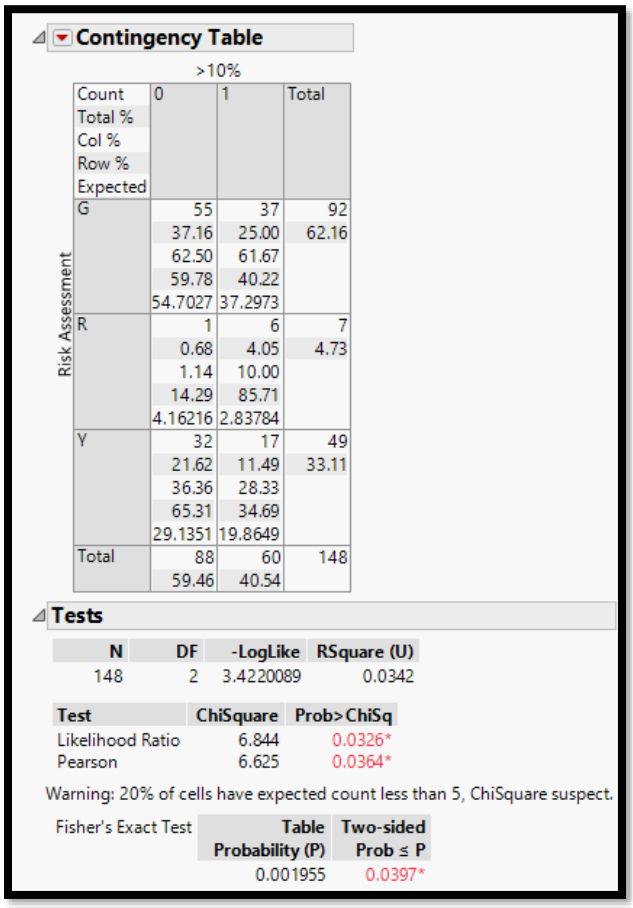
Requirements	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		64	44	108
		42.11	28.95	71.05
		71.11	70.97	
		59.26	40.74	
R		2	3	5
		1.32	1.97	3.29
		2.22	4.84	
		40.00	60.00	
Y		24	15	39
		15.79	9.87	25.66
		26.67	24.19	
		61.54	38.46	
Total		90	62	152
		59.21	40.79	

Tests

N	DF	-LogLike	RSquare (U)
152	2	0.41740807	0.0041

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	0.835	0.6588
Pearson	0.852	0.6532

Table D6 Contingency Table Results Development Risk Assessment Measure vs >10% Cost Variance



Contingency Table

Risk Assessment	>10%	0	1	Total
	Count	55	37	92
	Total %	37.16	25.00	62.16
	Col %	62.50	61.67	
	Row %	59.78	40.22	
	Expected	54.7027	37.2973	
R	Count	1	6	7
	Total %	0.68	4.05	4.73
	Col %	1.14	10.00	
	Row %	14.29	85.71	
	Expected	4.16216	2.83784	
Y	Count	32	17	49
	Total %	21.62	11.49	33.11
	Col %	36.36	28.33	
	Row %	65.31	34.69	
	Expected	29.1351	19.8649	
Total	Count	88	60	148
	Total %	59.46	40.54	

Tests

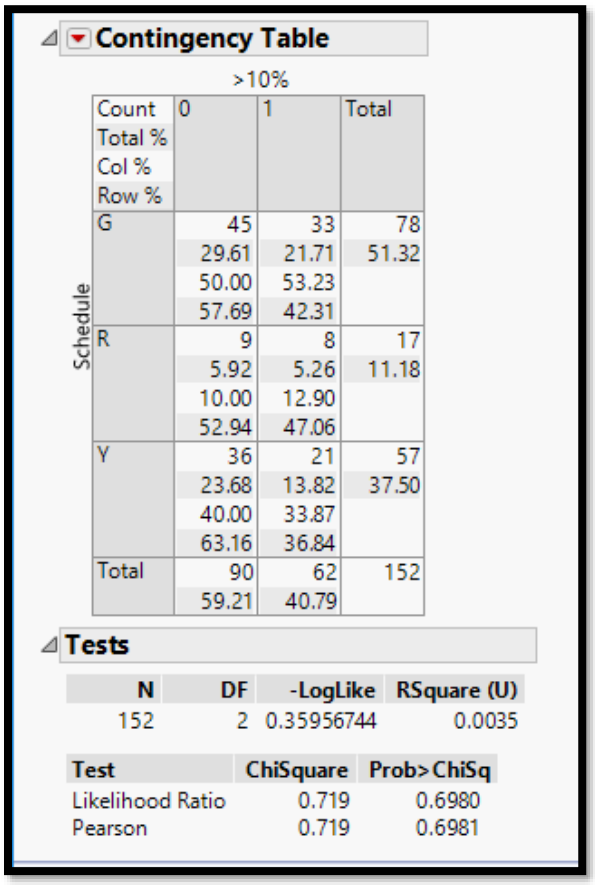
N	DF	-LogLike	RSquare (U)
148	2	3.4220089	0.0342

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.844	0.0326*
Pearson	6.625	0.0364*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.001955	0.0397*

Table D7 Contingency Table Results Development Schedule Measure vs >10% Cost Variance



Contingency Table

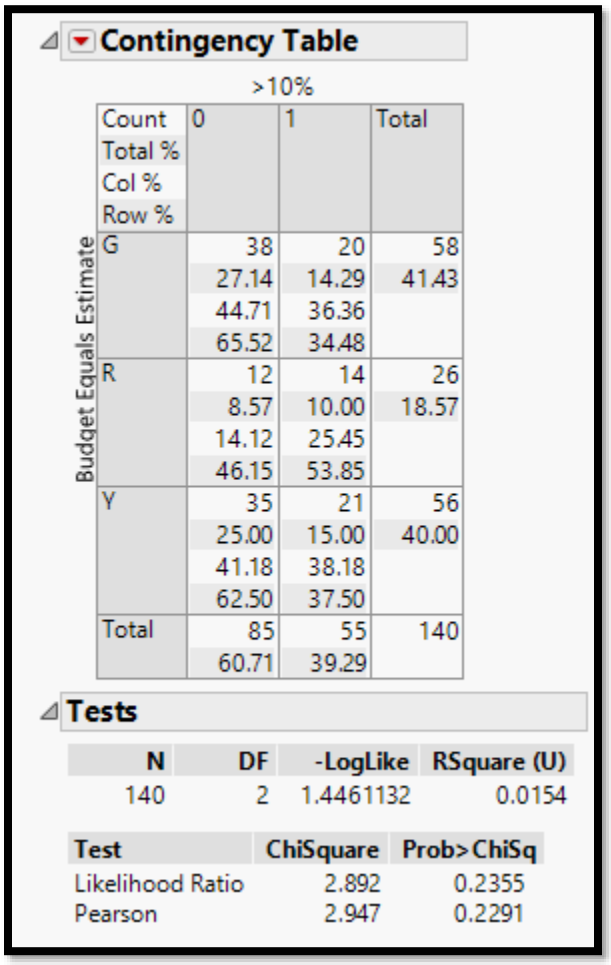
Schedule	>10%	Count	Total %	Col %	Row %	0	1	Total
G						45	33	78
						29.61	21.71	51.32
						50.00	53.23	
						57.69	42.31	
R						9	8	17
						5.92	5.26	11.18
						10.00	12.90	
						52.94	47.06	
Y						36	21	57
						23.68	13.82	37.50
						40.00	33.87	
						63.16	36.84	
Total						90	62	152
						59.21	40.79	

Tests

N	DF	-LogLike	RSquare (U)
152	2	0.35956744	0.0035

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	0.719	0.6980
Pearson	0.719	0.6981

Table D8 Contingency Table Results Procurement Budget Equal Estimate Measure vs >10% Cost Variance



Contingency Table

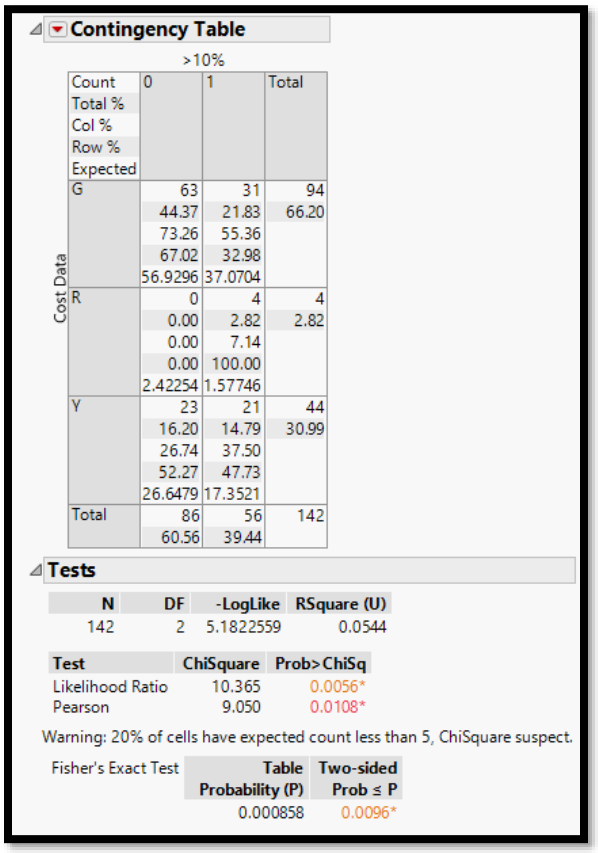
Budget Equals Estimate	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		38	20	58
		27.14	14.29	41.43
		44.71	36.36	
		65.52	34.48	
R		12	14	26
		8.57	10.00	18.57
		14.12	25.45	
		46.15	53.85	
Y		35	21	56
		25.00	15.00	40.00
		41.18	38.18	
		62.50	37.50	
Total		85	55	140
		60.71	39.29	

Tests

N	DF	-LogLike	RSquare (U)
140	2	1.4461132	0.0154

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	2.892	0.2355
Pearson	2.947	0.2291

Table D9 Contingency Table Results Procurement Cost Data Measure vs >10% Cost Variance



Contingency Table

Cost Data	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G	Count	63	31	94
	Total %	44.37	21.83	66.20
	Col %	73.26	55.36	
	Row %	67.02	32.98	
	Expected	56.9296	37.0704	
R	Count	0	4	4
	Total %	0.00	2.82	2.82
	Col %	0.00	7.14	
	Row %	0.00	100.00	
	Expected	2.42254	1.57746	
Y	Count	23	21	44
	Total %	16.20	14.79	30.99
	Col %	26.74	37.50	
	Row %	52.27	47.73	
	Expected	26.6479	17.3521	
Total	Count	86	56	142
	Total %	60.56	39.44	

Tests

N	DF	-LogLike	RSquare (U)
142	2	5.1822559	0.0544

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.365	0.0056*
Pearson	9.050	0.0108*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000858	0.0096*

Table D10 Contingency Table Results Procurement Crosschecks Measure vs >10% Cost Variance

Contingency Table

Crosschecks	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		47	21	68
		34.31	15.33	49.64
		57.32	38.18	
		69.12	30.88	
		40.7007	27.2993	
R		3	7	10
		2.19	5.11	7.30
		3.66	12.73	
		30.00	70.00	
		5.9854	4.0146	
Y		32	27	59
		23.36	19.71	43.07
		39.02	49.09	
		54.24	45.76	
		35.3139	23.6861	
Total		82	55	137
		59.85	40.15	

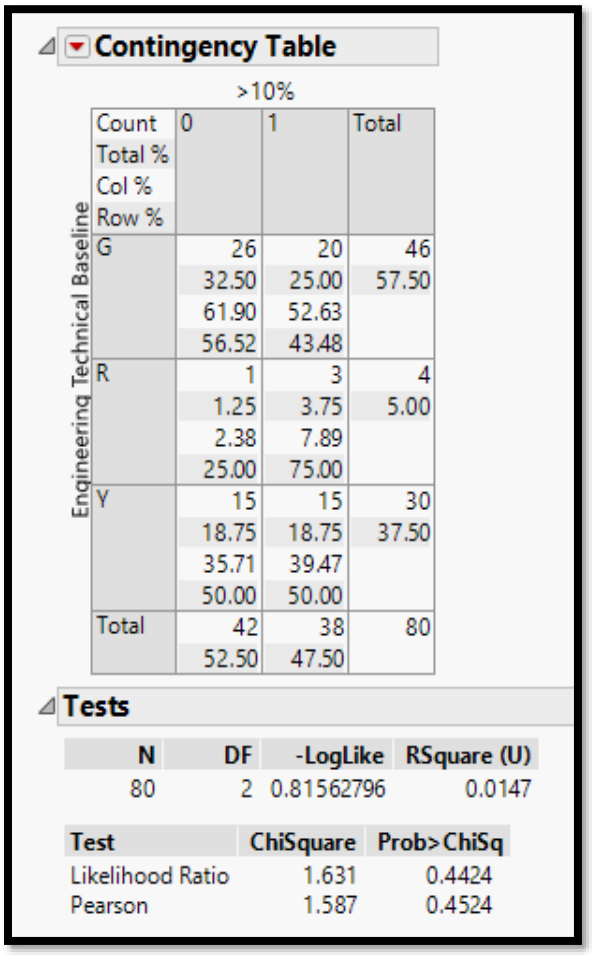
Tests

N	DF	-LogLike	RSquare (U)
137	2	3.4562600	0.0375

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.913	0.0315*
Pearson	6.912	0.0316*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.001313	0.0366*

Table D11 Contingency Table Results Procurement Engineering Technical Baseline Measure vs >10% Cost Variance



Contingency Table

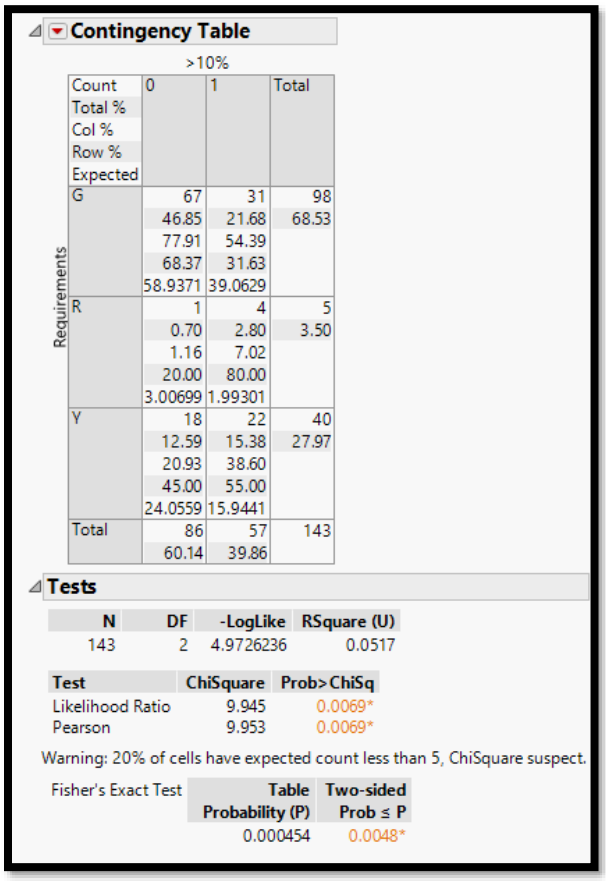
Engineering Technical Baseline	Count / Total % / Col % / Row %	>10% 0	>10% 1	Total
G		26	20	46
		32.50	25.00	57.50
		61.90	52.63	
		56.52	43.48	
R		1	3	4
		1.25	3.75	5.00
		2.38	7.89	
		25.00	75.00	
Y		15	15	30
		18.75	18.75	37.50
		35.71	39.47	
		50.00	50.00	
Total		42	38	80
		52.50	47.50	

Tests

N	DF	-LogLike	RSquare (U)
80	2	0.81562796	0.0147

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	1.631	0.4424
Pearson	1.587	0.4524

Table D12 Contingency Table Results Procurement Requirements Measure vs >10% Cost Variance



Contingency Table

Requirements	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		67	31	98
		46.85	21.68	68.53
		77.91	54.39	
		68.37	31.63	
		58.9371	39.0629	
R		1	4	5
		0.70	2.80	3.50
		1.16	7.02	
		20.00	80.00	
		3.00699	1.99301	
Y		18	22	40
		12.59	15.38	27.97
		20.93	38.60	
		45.00	55.00	
		24.0559	15.9441	
Total		86	57	143
		60.14	39.86	

Tests

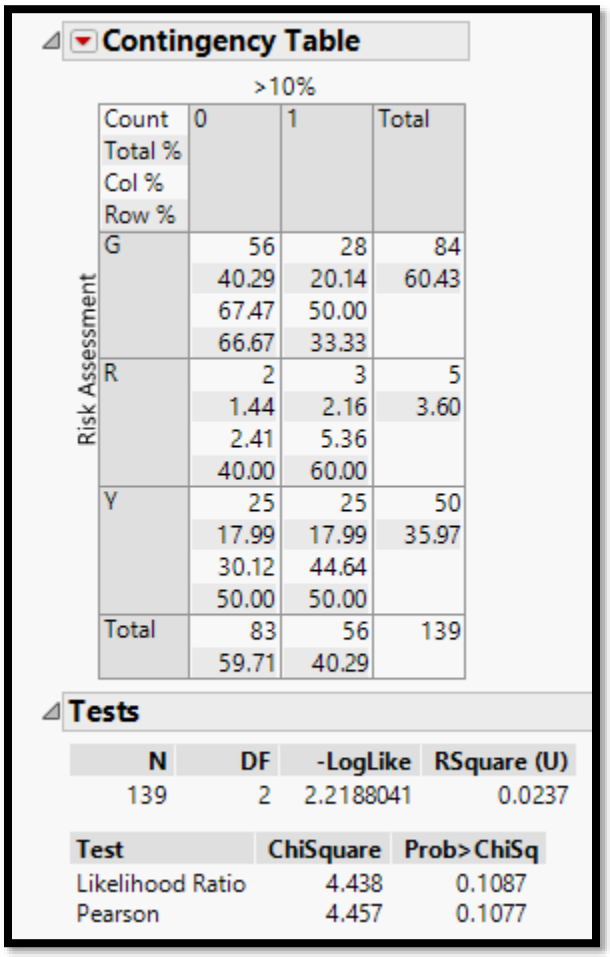
N	DF	-LogLike	RSquare (U)
143	2	4.9726236	0.0517

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	9.945	0.0069*
Pearson	9.953	0.0069*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	0.000454	0.0048*

Table D13 Contingency Table Results Procurement Risk Assessment Measure vs >10% Cost Variance



Contingency Table

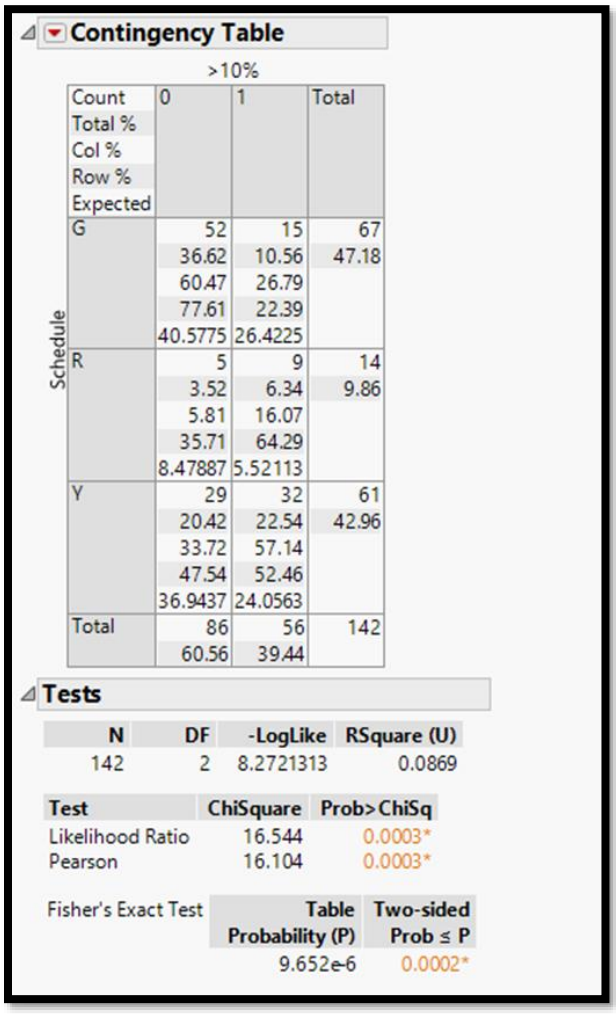
Risk Assessment	>10%	0	1	Total
	Count	56	28	84
	Total %	40.29	20.14	60.43
	Col %	67.47	50.00	
	Row %	66.67	33.33	
	R	2	3	5
		1.44	2.16	3.60
		2.41	5.36	
		40.00	60.00	
	Y	25	25	50
		17.99	17.99	35.97
		30.12	44.64	
		50.00	50.00	
	Total	83	56	139
		59.71	40.29	

Tests

N	DF	-LogLike	RSquare (U)
139	2	2.2188041	0.0237

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	4.438	0.1087
Pearson	4.457	0.1077

Table D14 Contingency Table Results Procurement Schedule Measure vs >10% Cost Variance



Contingency Table

Schedule	Count / Total % / Col % / Row % / Expected	>10% 0	>10% 1	Total
G		52	15	67
		36.62	10.56	47.18
		60.47	26.79	
		77.61	22.39	
		40.5775	26.4225	
R		5	9	14
		3.52	6.34	9.86
		5.81	16.07	
		35.71	64.29	
		8.47887	5.52113	
Y		29	32	61
		20.42	22.54	42.96
		33.72	57.14	
		47.54	52.46	
		36.9437	24.0563	
Total		86	56	142
		60.56	39.44	

Tests

N	DF	-LogLike	RSquare (U)
142	2	8.2721313	0.0869

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	16.544	0.0003*
Pearson	16.104	0.0003*

Fisher's Exact Test	Table Probability (P)	Two-sided Prob ≤ P
	9.652e-6	0.0002*

Bibliography

AcqNotes. (2019). *Acquisition Process: Exit & Entrance Criteria*. http://acqnotes.com/acqnote/acquisitions/exit-criteria

Arena, Mark A., Leonard, Robert S., Murray, Shelia E., Younossi, Obaid. (2006). *Historical Cost Growth of Completed Weapon System Programs*, Santa Monica, Calif, RAND Corporation.

Arena, M., Birkler, J., Blickstein, I., Nemfakos, C., Doll, A., Drezner, J., York, E. (2014). *Management Perspectives Pertaining to Root Cause Analyses of Nunn-McCurdy Breaches, Volume 6: Contractor Motivations and Anticipating Breaches*. RAND Corporation. Retrieved February 2, 2020, from www.jstor.org/stable/10.7249/j.ctt14bs3kg

Bartuseviciene, Ilona, and Evelina Sakalyte. (2013). *Organizational assessment: effectiveness vs. efficiency*. Social Transformations in Contemporary Society 1, no. 1: 45-53.

Bielecki, J. V., E.D.White. (2005). *Estimating Cost Growth For Schedule Changes: A Regression Approach.* Cost Engineering, 47 (8), 28-34.

Bolten, Joseph G., Robert S. Leonard, Mark V. Arena, Obaid Younossi, and Jerry M. Sollinger. (2008). *Sources of Weapon System Cost Growth: Analysis of 35 Major Defense Acquisition Programs*, Santa Monica, Calif.: RAND Corporation.

Brown, Gregory E., Edward D. White, Jonathan D. Ritschel & Michael J. Seibel. (2015) *Time Phasing Aircraft R&D Using the Weibull and Beta Distributions*. Journal of Cost Analysis and Parametrics, 8:3, 150-164. DOI:10.1080/1941658X.2015.1096219

Calcutt, Harry M. (1993). *Cost Growth in DOD Major Programs: A Historical Perspective*. The Industrial College of the Armed Forces, National Defense University Fort McNair, Washington DC.

Defense Acquisition University. (2019). *Defense Acquisition Guidebook*. Virginia: DAU.

Deneve, Allen J., Erin T. Ryan, Jonathan D. Ritschel, Christine Schubert Kabban. (2015). *Taming the Hurricane of Acquisition Cost Growth – Or At Least Predicting It.* Defense ARJ.

Drezner, Jeffrey A., Jeanne M. Jarvaise, Ronald Wayne Hess, Paul G. Hough, and D. Norton. (1993). *An Analysis of Weapon System Cost Growth*. Santa Monica, Calif.: RAND Corporation, MR-291-AF.

Dunning, David Chip Heath, Jerry M. Suls. (2004). *Flawed Self-Assessment: Implications for Health, Education, and the Workplace*. https://doi.org/10.1111/j.15291006.2004.00018.x

Foreman, James D. (2007). *Predicting the Effect of Longitudinal Variables on Cost and Schedule Performance.* MS thesis, AFIT/GIR/ENC/07M-01. Air Force Institute of Technology (AU), Wright-Patterson AFB OH.

Genest, D.C., E.D. White. (2005). *Predicting RDT&E Cost Growth, Journal of Cost Analysis and Management*. Fall, 1-12.

Government Accountability Office. (2009). *Cost estimating and assessment guide* (Report No. GAO-09-3SP). Washington, DC.

Government Accountability Office. (2010). *Defense Management: DOD Needs Better Information and Guidance to More Effectively Manage and Reduce Operating and Support Costs of Major Weapon Systems*. *Study*.

Government Accountability Office. (2012). *Defense Logistics: Improvements Needed to Enhance Oversight of Estimated Long- term Costs for Operating and Supporting Major Weapon Systems*. *Study*.

Government Accountability Office. (2018). *WEAPON SYSTEMS ANNUAL ASSESSMENT: Knowledge Gaps Pose Risks to Sustaining Recent Positive Trends.*

Harris, M.M., Schaubroeck, J. (1988). *A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings*. Personnel Psychology, 41, 43–62.

Jarvaise, Jeanne M., Jeffrey A. Drezner, and Daniel M. Norton. (1996). *The Defense System Cost Performance Database: Cost Growth Analysis Using Selected Acquisition Reports.* Santa Monica, CA: RAND Corporation. https://www.rand.org/pubs/monograph_reports/MR625.html

Kolar, D.W., Funder, D.C., Colvin, C.R. (1996). *Comparing the accuracy of personality judgments by the self and knowledgeable others.* Journal of Personality, 64, 331–337.

Kozlak, Scott J., White, Edward D., Ritschel, Jonathan D., Lucas, Brandon, & Seibel, Michael J. (2017). *Analyzing Cost Growth at Program Stages for DOD Aircraft*. Defense ARJ, Vol. 24 No. 3: 386-407, 2017.https://doi.org/10.22594/dau.16-763.24.03

Kozlak, Scott J. (2016). *Predicting Cost Growth Using Programs Reviews and Milestones for DoD Aircraft*. Theses and Dissertations. 246.

Keyser Donna J., Joan M. Lakoski, Sandraluz Lara-Cinisomo, Dana Schultz, Valerie L. Williams, Darlene F. Zellers, Harold Alan Pincus. (2008). *Advancing Institutional Efforts to Support Research Mentorship A Conceptual Framework and Self-Assessment Tool, Academic Medicine*. v. 83, no. 3, p. 217-225, RAND.

Lucas, Brandon M. (2004). *Creating Cost Growth Models for the Engineering and Manufacturing Development Phase of Acquisition Using Logistic and Multiple Regression*. MS thesis, AFIT/GCA/ENC/04-02. Air Force Institute of Technology (AU), Wright-Patterson AFB OH.

Lusthaus, Charles, Marie Helene Adrien, Gary Anderson and Fred Carden. (1999). *Enhancing organizational performance: a toolbox for self-assessment.* Publisher: IDRC 1999-01-01ISBN: 0889368708

McDaniel, C. J., E.D. White. (2007). *Predicting Engineering and Schedule Procurement Cost Growth for Major DoD Programs*. Journal of Public Procurement, 7 (3), 362-380.

Monaco, James V., Edward D. White III. (2005). *Investigating Schedule Slippage*. Defense Acquisition University, Alexandra, VA.

Moore, G.W., & White, E.D. (2005). *A regression approach for estimating procurement cost.* Journal of Public Procurement, 5 (2), 187-209.

Rossetti, M.B., E.D.White. (2004) *A Two Pronged Approach to Estimate Procurement Cost Growth in Major DoD Weapon Systems*. Journal of Cost Analysis and Management Winter, 11-21.

Rusnock, Christina F. (2008). *Predicting Cost and Schedule Growth for Military and Civil Space Systems*. MS Thesis, AFIT/GRD/ENC/08M-01, Air Force Institute of Technology, Wright Patterson AFB, OH (AU).

Schwartz, Moshe Charles V. O'Connor. (2016). *The Nunn-McCurdy Act: Background, Analysis, and Issues for Congress.* Congressional Research Service R41293, Washington, DC.

Searle, David A. (1999). *Twenty-five years of acquisition reform: where do we go from here?* Committee on Armed Services, House of Representatives, One Hundred Thirteenth Congress, first session.

Singer, Neil M. (1982). *Cost Growth in Weapon Systems: Recent Experience and Possible Remedies.* The Congress of the United States Congressional Budget Office, October.

Siow, Chong HR, Jian-Bo Yang, and B. G. Dale. (2001). *A new modeling framework for organizational self-assessment: development and application.* Quality Management Journal 8, no. 1: 34-47.

Stecher, Brian, and Sheila Nataraj Kirby. (2004). *Organizational Improvement and Accountability: Lessons for Education from Other Sectors*. 1st ed., RAND Corporation. JSTOR, www.jstor.org/stable/10.7249/mg136wfhf. Accessed 5 Feb. 2020

White, E.D., V.P. Sipple, M.A. Greiner. (2004). *Surveying Cost Growth*. Defense Acquisition Journal, 11 (1), 78-91.

Younossi, Obaid, Mark V. Arena, Robert S. Leonard, Charles Robert Roll, Jr., Arvind Jain, and Jerry M. Sollinger. (2007). *Is Weapon System Cost Getting Better or Worse?* Santa Monica, CA: RAND Corporation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
10-02-2020	Master's Thesis	Sept 2019 – March 2020

TITLE AND SUBTITLE

The Utility of Self-Assessment in Predicting Program Office Estimate Accuracy

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Luketic, Dana P., MSgt, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/ENV)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENV-MS-20-M-225

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AIR FORCE LIFECYCLE MANAGEMENT CENTER
1865 Fourth St, Bldg 14, WPAFB, OH 45344
937-656-5504 and shawn.valentine@us.af.mil
ATTN: Shawn Valentine

10. SPONSOR/MONITOR'S ACRONYM(S)

AFLCMC/FZC

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
The ability of the Program Offices to provide accurate cost estimates is an essential element in planning and programming. Historically, cost estimating has led to budget overruns and continues to be an area of scrutiny and concern. A series of legislative reforms have sought to address each of these perceived underlying causes which are located at all levels of decision making – from the SPO to CADE. The current study is specifically interested in determining how well SPOs are doing. There have not been comprehensive studies on SPO performance. In large part, this deficiency is due to the inability to systematically assess the SPOs. However, a new consolidation of data by AFLCMC has recently made it possible to do such a study. The AFLCMC's program office estimates in this study will look at the SPOs of AFLCMC and evaluate their cost estimates for growth and also determine if their established method of self-assessment provides a predictor of the overall future accuracy of the program estimate.

15. SUBJECT TERMS
Cost Estimates, Cost Growth, Program Office, Self-Assessment

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Lt Col Scott Drylie, AFIT/ENV
U	U	U	UU	97	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-3636 x4441; Scott.drylie@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18