9-1998

An Investigation of the Reliability of the Contractor Performance Assessment Reporting System (CPARS)

John C. Odum

AN INVESTIGATION OF THE RELIABILITY
OF THE CONTRACTOR PERFORMANCE
ASSESSMENT REPORTING SYSTEM (CPARS)

THESIS

John C. Odum, Captain, USAF

AFIT/GCA/LAS/98S-6

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GCA/LAS/98S-6

AN INVESTIGATION OF THE RELIABILITY OF THE CONTRACTOR PERFORMANCE ASSESSMENT REPORTING SYSTEM (CPARS)

THESIS

John C. Odum, Captain, USAF

AFIT/GCA/LAS/98S-6

AFIT/GCA/LAS/98S-6

AN INVESTIGATION OF THE RELIABILITY OF THE CONTRACTOR PERFORMANCE ASSESSMENT REPORTING SYSTEM (CPARS)

THESIS

Presented to the Faculty of the Graduate School of Logistics

and Acquisition Management of the 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

John C. Odum, B.S.

Captain, USAF

September 1998

Acknowledgments

This thesis would not have been possible without the support of some very key and influential people. I would like to take this opportunity to say Thank You.

First and foremost, I would like to thank my thesis committee: my advisors, Lt Col Steve Giuliano (until his retirement) and Maj Caisson Vickery (after Lt Col Giuliano's retirement), and my reader, Maj Daryl Hauck. Thank you for your support and guidance. Tom Fowler, Bonnie Taylor, Nancy Barlow and Bob Nolin from my sponsoring organization, ASC/SYG, and Col Avery Sledge, Donna Hatfield, Roger Hanson, and Mike Squires from AFMC/PK have been instrumental in this effort. Also, Nancy Wiviott kept me motivated and smiling through her reassuring pressure tactics. Again, without their collective support, this thesis would not have happened.

I would also like to thank several fellow students and their families: Dave Bach, Kevin Brummert, Sharon Gibson, John "G" Glover, Brian Kehl, Brad McDonald, Phil Mischler, Jimmy "Seeker" Sikra, Lance Whitfill, and Mike Wilson. Your friendship and support made my AFIT experience exactly that, an experience.

Most importantly, I want to thank my children, Chris and Brittney, for their patience and love during this assiduous time. I look forward to "paying you back" and spending the rest of my summers with you both.

John Odum

Table of Contents

Page

Page

Page

List of Figures

Figure Page

List of Tables

Table | Page

AFIT/GCA/LAS/98S-6

Abstract

An Industry trend is to establish long-term relationships with reliable suppliers. One of the criteria used to pick these "reliable suppliers" is past performance. The Department of Defense is also attempting to capitalize on this logical trend to the maximum extent possible by using past performance as an evaluation factor in source selections. Air Force Material Command (AFMC) employs the Contractor Performance Assessment Reporting System (CPARS). This thesis examines the reliability of the CPARS.

This study began with 149 records from the Aeronautical Systems Center CPARS database. The evaluation relied on three basic techniques: correlation tests, a Tukey multiple comparison procedure, and linear regression.

This thesis found, despite the fact that policy mandates color ratings be based on period objective measures, the cost color ratings were more consistent with cumulative objective measures. Even so, the strength of this relationship has degraded significantly over time. With respect to schedule, the reliability is improving significantly, but period objective measures are not yet significantly correlated with schedule color ratings. The author recommends that AFMC either change CPARS cost rating policy to reflect the use of cumulative objective measures or provide additional training so evaluators better understand what is assessed during a CPARS rating period.

AN INVESTIGATION OF THE RELIABILITY

OF THE CONTRACTOR PERFORMANCE

ASSESSMENT REPORTING SYSTEM (CPARS)

I. Introduction

General Issue

As a reaction to increased global competition and technological innovations an Industry trend over the past few decades has been to establish long-term relationships with fewer, more reliable suppliers (Spekman, 1988:75, Little, 1996:5). Traditionally, buyers simply used an adversarial model to "minimize the price of purchased goods and services." This adversarial approach assumed "that there are no differences in suppliers' abilities to provide value-added services, technology gains, process innovations, and other means of gaining differential advantage" (Spekman, 1988:76). Now, the "standard criteria of quality, price, and delivery are necessary-but-not-sufficient conditions for consideration" (Spekman, 1988:79). In fact, most buyers now realize that not all suppliers make good partners and past performance must be used as an evaluation criteria (Spekman, 1988:80). This understanding has driven both buyers and suppliers to establish long-term relationships with one another. An important factor in this process of selecting reliable suppliers has been evaluating the performance of particular suppliers based on a number of criteria, including past performance (Little, 1996:5).

The Department of Defense (DoD) also wants to capitalize on this logical trend. While the DoD began using PPI sporadically as early as 1961, an Office of Federal

Procurement Policy (OFPP) 1994 pilot program catapulted its importance in source selections. Under this pilot program, 20 federal agencies will be awarding 61 future contracts using past performance as a major evaluation factor (Ichniowski, 1994:11). The latest guidance is from a 20 Nov 1997 memorandum to the Services signed by Undersecretary of Defense for Acquisition and Technology, Dr. Jacques S. Gansler. The Gansler memo stated, “Collection of Past Performance Information (PPI) is critical to using this information to obtain best value goods and services” (Gansler, 1997:1). Air Force Material Command (AFMC) utilizes the Contractor Performance Assessment Reporting System (CPARS) to collect past performance data. “The sole purpose of the CPARS is to ensure a commandwide data base of contractor performance information is current and available for use in responsibility determinations and in formal and informal source selections” (AFMC, 1997:1).

Background

As stated earlier, one primary driver for establishing long-term relationships with reliable suppliers is that global competition has increased. Increased global competition pressures corporations to choose good partners in order to “achieve a stronger competitive position in the marketplace” (Spekman, 1988:75). Through policies such as the Competition in Contracting Act (CICA) of 1984, the DoD has been required to promote competition, albeit in an adversarial type fashion. However, in an effort to obtain the best partners, albeit "adversarial partners", the Air Force (AF) has increased the use of PPI during Source Selection and the use of Integrated Product Teams (IPTs) once the contract has been awarded.

Another major reason for the DoD to use PPI to select the best value suppliers lies in the fact that the DoD is facing massive budget cutbacks. The budget reductions are affecting not only total defense spending but also the future of defense, of which Research and Development (R&D) is a major player. According to the U.S. Office of Management and Budget (OMB), the total DoD budget for FY98 will be $254.9B. This represents a non-inflation-adjusted reduction of nearly 20% from the total FY89 DoD budget of $303.6B (U.S. Bureau, 1996:351). The National Science Foundation also indicates that the R&D Budget Authority (BA) has decreased nearly every year from FY90 to FY98 in constant FY92 terms. The total R&D BA has dropped from $42.8M (BY92$) in FY90 to $33.4M (BY92$), or a 22% reduction, in the FY98 proposal (NSF, 1997:54). These draconian budget cutbacks solidify the need to select the "best value" vendors during Source Selections. For several years DoD has recognized that "Focusing on past accomplishment provides a powerful incentive for improvement in these difficult fiscal times" (Weidenbaum, 1992:51). Finally, because of dwindling resources, a cost-effective means for collecting the information to be used in source selections proves to be a critical issue DoD faces in implementing past performance policy (Little, 1996:16).

In a report to the Office of the Secretary of Defense, Arthur D. Little, Inc., stated further reasons to use PPI. These reasons were: 1) Using PPI makes good business sense; 2) Using PPI is currently being used successfully (on a limited scale); and 3) Using PPI can be tailored to fit specific circumstances (Little, 1996:84). However, this emphasis on past performance may cause distress for both Industry and the Government. A major Industry concern would ask, "Does performing 'at all costs' to keep high 'perceived arbitrary' past performance ratings significantly diminish profit potential?" From the Government perspective, the "best value" suppliers are desired whenever possible. If DoD could avoid much of the meticulous source selection process, a great deal of cost and schedule savings could be realized. In short, superior performance on current contracts should reward contractors with additional future contracts. The ratings used must be reliable so that excellent performance ratings on previous contracts should predict excellent future performance.

The 1994 OFPP study provides an example of the benefits of using PPI. "In 1994, 25 civilian agencies, the military services and the Defense Logistics Agency pledged to conduct pilot tests of the idea of using past performance data" (Laurent, 1997:23). The OFPP study "showed that on 30 contracts re-competed using past performance information, the average customer satisfaction level increased 21 percent over the previous contract" (Laurent, 1997:23). In addition, OFPP Administrator Steven Kelman reported that contractors are working harder on government contracts than in the past because they want "good report cards." However, "Kelman's biggest worry is that source selection officials will inflate vendor's grades and fail to discriminate between bidders whose past performance scores are very close" (Laurent, 1997:23-24).
The Contractor, on the other hand, obviously does not want to be punished for weaker cost and schedule performance while providing an excellent technical or a high quality product. "Past performance ratings for government contracts have long been criticized as

inconsistent, subjective and poorly organized" (DoD sets Contractors Standards, 1997:10). Donna Ireton, contracts director for Advanced Systems Development, Inc., argues that, "There is no standardized approach, no centralized database" (Burman, 1997:60). Other criticisms of using past performance have been that it represents another barrier to entry, and evaluations tend to be inflated (Burman, 1997:60).

In response to some of the above complaints, the DoD has recently attempted to improve the way it rates contractors on their performance. In the memorandum previously mentioned, Dr. Gansler "has established a five-level past performance rating system for almost all categories and sectors of contracts" (DoD Sets Contractor Standards, 1998:10). This new policy marks the "first large-scale attempt at standardizing the collection of past performance information." Also, the DoD is developing automation capability to view the PPI records on-line (DoD Sets Contractor Standards, 1998:10).

Market trends are forcing corporations to change their buyer-supplier relationships from arms-length adversarial relationships to nearly partnerships. The USAF, however, cannot realize the full benefits from long-term relationships with a small set of suppliers to the same potential as in Industry due to necessary socioeconomic factors. Some socioeconomic factors include small business firms, disadvantaged or minority-owned businesses, and labor surplus areas. Other related factors are concerned with maintaining a strong industrial base with a surge capability while maintaining a leading edge in defense technology. Yet, the DoD wants to follow this market trend to the maximum extent possible. Although much research must be put into developing ratings that actually incentivize good performance, the foundation has been laid. Of course, the DoD will determine the strength of this foundation only by using it. "'If we do it right we'll get contractors to perform above satisfactory,' says David Drabkin, assistant deputy undersecretary of Defense for acquisition process and policies. 'They'll improve their performance today'" (DoD Sets Contractor Standards, 1998:10).

Problem Statement and Investigative Questions

According to the Arthur D. Little, Inc. report, "CPARS very consistently performs its intended purpose" (Little, 1996:20). However, the Little report does not address the reliability or the validity of the CPARS process. Because, "Measurements can be reliable without being valid for a stated purpose, it is impossible for a measurement system to be valid without being reliable" (Kachigan, 1991:141). Therefore, determining the actual reliability of the CPARS database is the first step in ensuring the validity of the CPARS. The Problem Statement to be answered then by this thesis will be to "Investigate the reliability of the CPARS." By finding the results of three specific investigative questions, the problem statement can be answered:

1. The first question examines the reliability of the CPARS ratings. Do the best performances always receive the best ratings? In terms of this study, do objective performance measures positively correlate with performance ratings?
2. The second question examines the reliability of the CPAR ratings over time. More specifically, how has the reliability of the CPAR ratings changed over time? Has the reliability increased, decreased, or stayed the same?

3. The final question explores the relationship between CPARS and figures of profitability such as Return on Equity (ROE) percentage and Return on Investment (ROI) percentage. This question will help evaluate whether the DoD policy rewards good contractor performance. Specifically, do performance criteria and/or ratings positively correlate with performance ratings?

Scope of the Study

The principal statistical analysis used in this study tests for differences between the correlation coefficients of the Cost and Schedule Control factor color ratings and actual performance given that period. For parts of the questions above, a Spearman's rank correlation coefficient will help determine the degree of correlation between the color ratings and actual performances. The procedures used to measure and compare the coefficients are discussed in Chapter III. A Tukey multiple comparison procedure will determine if the objective measurements (cost and schedule variances) are different between the color ratings (Devore, 1991:381). Regression will also be performed for the trend analysis.

The method applied in this study relies on actual Cost-type, Aeronautical System Center (ASC) contracts from the AFMC CPARS database. Specific cost data has been obtained from the ASC Cost Library, System Program Offices, and the Internet. Some of the data used in this study has been masked in order to avoid inadvertent release of proprietary information. Profitability data has been obtained from the Internet.

Organization of the Study

The next chapter provides an historical look at DoD use of PPI. The chapter then describes how a Corporation's performance measures as a DoD supplier tie to the CPARS through certain objective contract performance criteria. Finally, Chapter II describes the CPARS process as well as the requirements specified in the Gansler memo. The methodology that will be used to test for correlations, the differences between mean color rating variance percentages and regression tests is detailed in the third chapter. Chapter III also contains the limitations of this study. The data analysis and findings comprise Chapter IV. The final chapter discusses the conclusions and recommendations for further research in this area of study.

II. Literature Review

Chapter Overview

The following section provides a historical review of Government use of Contractor PPI. Also included in the first section is a model displaying how performance measures of both corporation and contract performance flow through the CPARS to provide ratings for Source Selection Evaluations and feedback to the contractor. The next discussion encompasses how DoD measures PPI and illustrates both DoD and Industry concerns with the entire process. The third section describes the CPARS and the collection process in conjunction with the requirements specified by the Gansler memo. The final section summarizes this chapter and discusses propositions that relate to the investigative questions listed in Chapter I.

Contractor Past Performance Information

Historical use of PPI

Attention to past performance in the DoD acquisition community has increased significantly over the past few years. In 1995, the Deputy Under Secretary of Defense (Acquisition Reform) contracted with Arthur D. Little, Inc. to study the Department's past performance systems and to develop a proposal for implementation of a Department-wide process for the effective use and collection of past performance information (DoDAR Website, 1998). The resulting study currently sets the benchmark by which PPI systems are evaluated. Specifically, Little recommends that a PPI system possess the qualities listed in Table 1.

Table 1. Qualities for an Effective PPI System (Little, 1996:86-87)

Implements decentralized approach with general guidelines
Focuses on similar product areas or services
Views total program context
Horizontally integrated through business area alliances
User helps define what gets collected and when
Easy to understand and explain
Information shared among organizations

Little defines PPI as "relevant information regarding a contractor's actions under previously awarded contracts" (Little, 1996:12). This definition of PPI includes the contractor's record of conforming to specifications, forecasting and containing costs, adhering to contract schedules, establishing a commitment to customer satisfaction, and maintaining a business-like concern for their customer (Little, 1996:12).

According to Brislawn and Dowd, PPI consists of Agency evaluations, CPARS or other rating systems, federal, state, and local government as well as other private contracts identified in the contractor's proposal, contractor self-assessments, user and buyer evaluations, and performance qualifications (Brislawn and Dowd, 1996:18). Brislawn and Dowd further contend that, "The greater the amount of relevant information considered, the more accurate the evaluation of the contractor's past performance and the more accurate the assessment of the contractor's ability to perform the proposed contract" (Brislawn and Dowd, 1996:18).

Many Industrial firms have abandoned arm-length relationships with their suppliers in favor of friendly, long-term relationships (Han, Wilson, & Dant, 1993:337). In fact, global competition "has made American companies aware of the importance of having close relationships not only with their customers, but also with their suppliers" (Han, Wilson, & Dant, 1993:331). Also, most companies are "consciously making an effort to reduce their supplier base" (Han, Wilson, & Dant, 1993:337). Larson and Kulchitsky add, "The evidence is compelling – single sourcing and supplier certification have favorable impacts on buyer/supplier relationships" (Larson and Kulchitsky, 1998:80). When selecting the appropriate supplier, companies must use some measure of performance. Timmerman suggests that, "the most important indicator of a supplier's ability to add value to a transaction is usually its record of performance in previous transactions" (Timmerman, 1986:2). Similarly, "private companies and consumers routinely return to vendors who prove their worth" (Ichniowski with Rubin, 1994:83). Further, "Past performance complements the contractor's understanding of contract requirements (as described in the proposal) with a measure of their *actual* ability to perform" (Brislawn and Dowd, 1996:16). Thus, it is prudent to use PPI as an indicator of future performance during DoD source selections.

DoD has used PPI periodically over the last 30 years. Each DoD PPI initiative, however, has been abandoned because the perceived benefits have not outweighed the cost and administrative burden (Little, 1996:5). Further, when PPI has been used in source selection evaluations, the data has only been gathered on an ad hoc basis (Little, 1996:4).

Table 2 provides a timeline of PPI use by the Government. The use of PPI began in 1961 when President Kennedy appointed the Bell Committee. This initial use of PPI was cancelled in 1970 because it was deemed costly and ineffective (Sumpter, 1998:2). PPI gained more importance with the 1986 Packard Commission Report, which stated DoD should make greater use of commercial-style practices. In particular, the DoD could reduce costs by maintaining a list of qualified suppliers that have held high standards of product quality and reliability (President's, 1986:62-63). In January of 1993, OFPP issued past performance policy through Policy Letter 92-5. This letter required that past performance be a mandatory evaluation factor in competitive negotiations (Scott, 1995:4).

Recent emphasis on using PPI has been provided by passage of the Federal Acquisition Streamlining Act (FASA) of 1994 and the FAR 15 rewrite for the use of PPI (Sumpter, 1998:3). The FASA of 1994 allowed government source selections to behave more like industry source selections by, "requiring a comparative assessment of contractors' past performance in the source selection process" (Brislawn and Dowd, 1996:16). In May 1995, SAF/AQ released the first 8 of 11 Lightning Bolt Initiatives (LBI) (Dept of the Air Force, 1996:35). The intent of LBI #6, "Enhance the role of past performance in source selections" was to change the CPARS to "collect accurate, comprehensive evaluations of contractors and subcontractors" (Dept of the Air Force, 1996:39).

Table 2. Government Use of Contractor PPI (Sumpter, 1998:2-3)

1961	President Kennedy appointed the Bell Committee, a "Blue Ribbon" committee that recommended an exchange of information between agencies regarding contractor evaluations
1962	President directed an elaborate Contractor Performance Evaluation (CPE) system be devised
1970	President's Blue Ribbon Defense Panel cancelled CPE as costly and ineffective
1978	Air Force initiates a field test at four product divisions to test effectiveness of evaluating past performance
1981	Use of past performance without reliance on a formal system in source selections was one of the 32 Carlucci Initiatives
1984	Air Force test discontinued based on consensus that PPI collection must be efficient and include data from buying commands as well as administration officials
1984	The Competition in Contracting Act was passed advocating the use of past performance
1986	President Reagan's Packard commission recommended that law and regulation should include increased use of commercial style competition emphasizing quality and established performance as well as price
1987	Air Force conducts Project STAR study that concluded use of PPI was ineffective because it was inconsistent and thus unreliable
1988	Air Force initiated the Contractor Performance Assessment Reporting System (CPARS) as a command wide performance data base
1989	Secretary of Defense, Dick Cheney, chartered a joint OSD-DoD task force to expand the CPARS concept DoD wide that concluded a DoD-wide system was not feasible
1993	The Office of Federal Procurement Policy (OFPP) issued Policy Letter 92-5 requiring the executive agencies to collect and use past performance information
1994	The Federal Acquisition Streamlining Act (FASA) signed into law
1995	FAR coverage and the OFPP Draft Best Practices Guide on Past Performance published
1995	USD(A&T) approved a study contract that recommended collection of PPI by business sector
1995	DFAR coverage was drafted
1996	Air Force and Navy Aeronautical sector develops a joint CPARS format
1996	DFAR case was withdrawn due to lack of consensus on methodology among the components
1997	USD(A&T) issues new policy on collection of PPI and the FAR 15 rewrite team generates new guidance for the use of PPI

Current use of PPI

Currently, DoD uses three types of PPI systems. Each of these systems is defined as, "an ongoing effort to collect and record past performance information for subsequent use in determining contractor eligibility and selection" (Little, 1996:14). The three types of systems are 1) Performance appraisal systems, 2) Performance tracking systems, and 3) Performance certification systems (Little, 1996:14). Figure 1 shows that CPARS is one of many existing systems within each of the three categories.

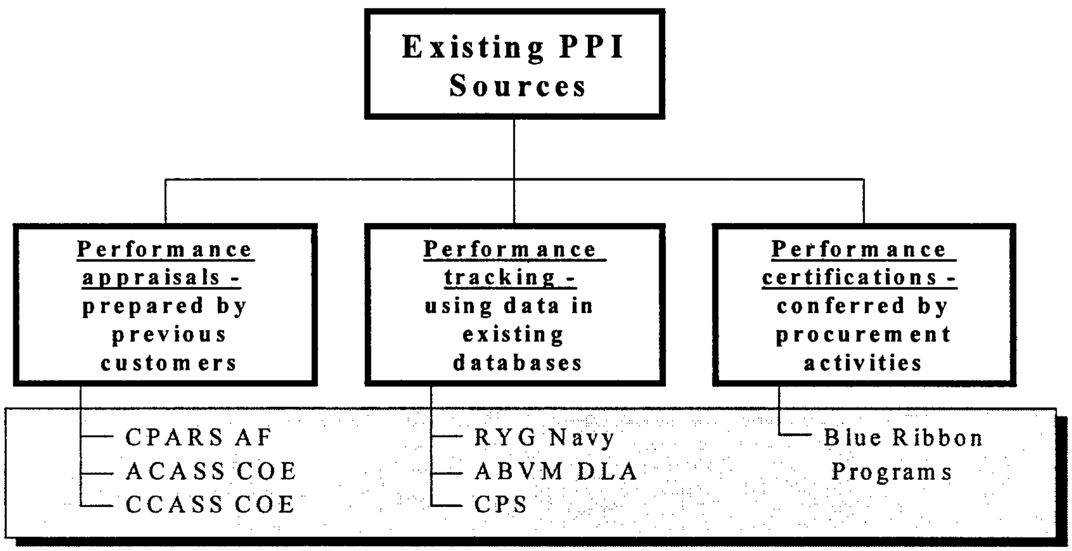

Figure 1. PPI Systems Used Within DoD (Little, 1996:14)

Since these systems were developed by different agencies for different purposes, they provide different utility. For example, a main difference between performance appraisal systems and performance tracking systems is the number of factors being evaluated. Performance appraisal systems cover eleven or more factors whereas the performance tracking systems usually assess only two or three factors (Little, 1996:14).

Past Performance Information Dissemination Model

Figure 2 illustrates the flow of information through an evaluation system such as the CPARS. Corporations use different performance criteria, or benchmarks, to determine how they measure up to industry leaders. Measurements such as Return on Equity (ROE) or Return on Investment (ROI) aid in providing a measure of the profitability of a corporation at the aggregate level. The performance of the corporation is an amalgamation of that company's performance on each of the contracts. Depending upon the type of effort, different criteria can be used to determine contract performance. For DoD Cost-reimbursement contracts, cost and schedule variances are measured using an Earned Value Management System (EVMS) and reported on Cost Performance Reports (CPR) or Cost/Schedule Status Reports (C/SSR). Other measurements such as management capability, technical quality, or other appropriate factors can be evaluated and reported through other vehicles. These measurements form the basis of the CPARS ratings. Because subjectivity is involved when determining overall performance in critical areas, subjective differences and rater bias may cloud the rating assigned to describe the contractor's performance. Allowing contractors an opportunity to respond to ratings combined with the standardization discussed in the Common DoD Assessment Rating System section is intended to minimize subjective influences and rater bias. This data is then collected and stored in the CPARS database.

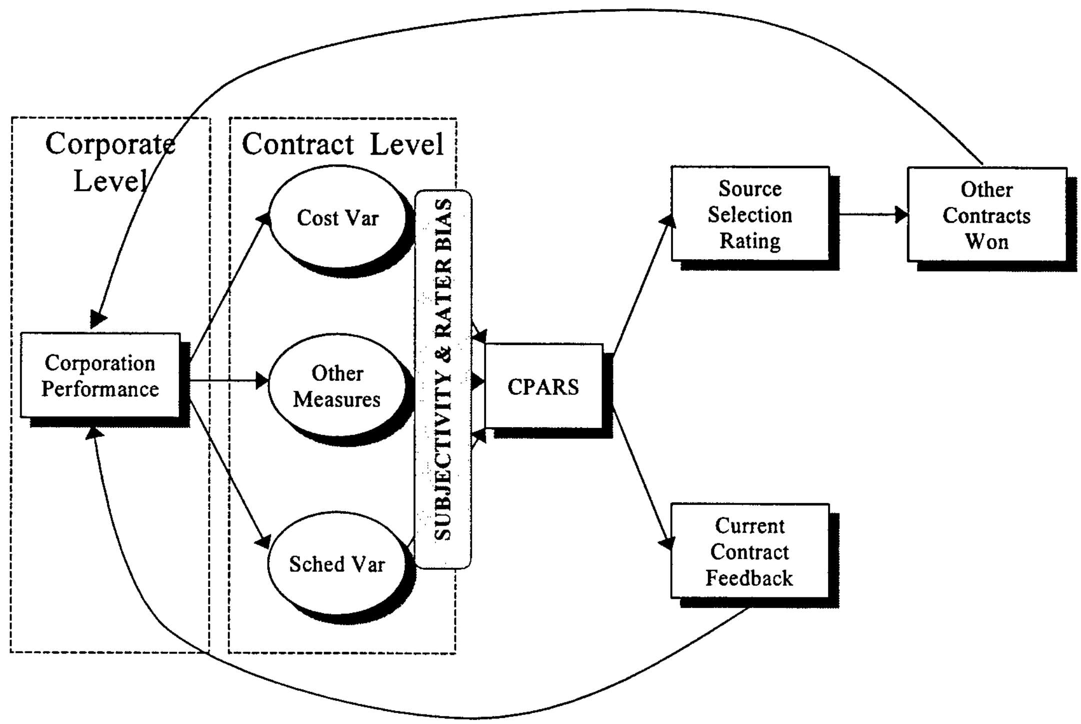

Figure 2. Past Performance Information Dissemination Model

As Figure 2 depicts, two uses of the CPAR information are providing information to source selections and feedback on current efforts. The use of CPARS information in source selections will be discussed in the CPARS section. "Feedback to suppliers" the Little report states, "is a very important ingredient in an effective supplier evaluation program. This provides needed information on quality to suppliers for their own improvement processes" (Little, 1996:31). This feedback is critical to any supplier evaluation program because it enables both the Contractor and Government the "opportunity to improve the product, reduce costs, and improve service" (Little, 1996:32). The feedback then helps the corporation improve its performance.

Contractor Performance Assessment Reporting System (CPARS)

Synopsis

According to AFMCI 64-107, the CPARS is a semi-automated AFMC database which ensures that contractor performance information is current and available for use in responsibility determinations in formal and informal source selections. The CPARS' intention is to efficiently communicate contractor past performance to source selection officials (AFMC, 1997: 1).

The CPARS evaluates both positive and negative performance on a given contract during a specific time interval. An initial report is required for new contracts meeting certain thresholds discussed in the Business Sector section. The initial report evaluates performance on at least the first 180 days of the contract, but no more than the first 365 days of the contract. Intermediate reports are then required every twelve months throughout the period of performance of the contract. The intermediate reports must discuss only the performance since the preceding CPAR. The final report is "completed upon contract termination, transfer of program management responsibility outside of AFMC, or the delivery of the final major end item on contract or completion of the period of performance" (AFMC, 1997:5). Out-of-cycle reports must be completed as needed (AFMC, 1997:4-5).

Each report must be based on objective, supportable facts. Although subjective assessments should be provided, the evaluation should not contain speculation. The CPARS allows the contractors opportunities to respond to program manager comments, which facilitates objective and consistent evaluations. Finally, summary data can be used to evaluate industry performance provided that the data does not reveal specific contract or contractor performance in any form (AFMC, 1997:1-2).

Business Sectors

The attachment to Dr. Jacques S. Gansler's 20 Nov 1997 Memorandum to the services defined two main business sectors that encompass DoD acquisition. These two sectors are Key Business Sectors and Unique Business Sectors (Gansler, 1997:3). The Key Business Sector is divided into four subsectors: Systems, Services, Operations Support, and Information Technology (Gansler, 1997:3-5). Likewise, the Unique Business Sector includes Construction and Architect-Engineering, Health Care, Fuels, and Science and Technology acquisitions (Gansler, 1997:6-7). This division into sectors is consistent with the Little report findings that, "Although the industry programs varied in many of their details, one of the common elements was a recognition that successful program needed to be tailored to **discrete** business areas" (Little, 1996:5). Figure 3 provides a graphical representation of these sectors.

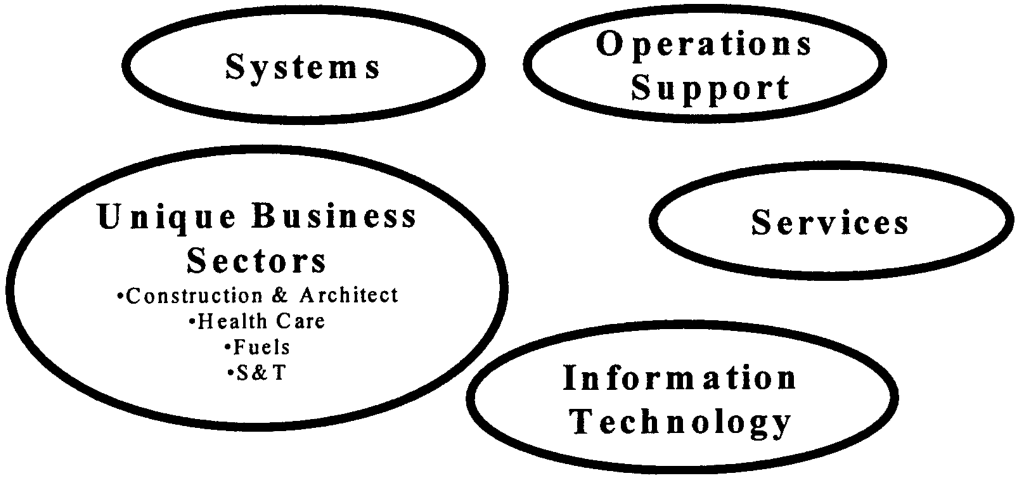

Figure 3. Key Business Sectors (Sledge, 1998:6)

Gansler's memo also specifies thresholds when PPI will be collected and which elements will be evaluated for each sector. For the Systems sector, the threshold is $5,000,000 or more and the assessment elements are Technical, Schedule, Cost Control, and Management. The other three Key Business sectors, Services, Operational Support, and Information Technology, all evaluate the same assessment elements but have different dollar thresholds (Gansler, 1997:7). Table 3 lists the different acquisitions within each of the Key Business sectors as well as the respective thresholds and assessment elements.

Table 3. Key Business Sectors (Gansler, 1997:3-8)

Sector	Acquisitions	Thresholds	Assessment Elements
Systems	Aircraft Shipbuilding Space Ordnance Ground Vehicles Training Systems Other Systems	≥ $5 M	Technical Schedule Cost Control Management
Services	Professional/Technical & Management Support Repair & Overhaul Installation Services	≥ $1 M	Quality of Product/Service Schedule Cost Control Business Relations Management of Key Personnel
Operational Support	Mechanical Structural Electronics Electrical Ammunition Troop Support Base Supplies	≥ $5 M	Quality of Product/Service Schedule Cost Control Business Relations
Information Technology	Software Hardware Telecommunications Equipment or Services	≥ $1 M	Quality of Product/Service Schedule Cost Control Business Relations Management of Key Personnel

Common DoD Assessment Rating System

Dr. Gansler's memo defined five categories of ratings for use in all acquisitions except Construction and Architect-Engineering (Gansler, 1997:9). The CPARS then was required to expand from four to five rating elements. In an 11 Aug 1997 Memorandum, Mr. R. Noel Longuemare provided two reasons for the DoD adoption of a five-point system. First, smaller program offices "tend to have fewer personnel and less time to provide the kind of narrative evaluation that is necessary to the successful operation of a four-point system" (Longuemare, 1997). The second reason is that the fifth element will help "the source selection authority to distinguish between offerors in deciding best value to the government" (Longuemare, 1997). Table 4 summarizes these five categories, their definitions, and how they correspond to the CPARS color ratings

Table 4. Summary of Ratings

DoD Category	Definition	Color Rating
Exceptional	Performance meets contractual requirements and exceeds many – corrective actions were highly effective	Blue
Very Good	Performance meets contractual requirements and exceeds some – corrective actions were effective	Purple
Satisfactory	Performance meets contractual requirements – corrective actions were satisfactory	Green
Marginal	Performance does not meet some contractual requirements – corrective actions were marginally effective or not implemented	Yellow
Unsatisfactory	Performance does not meet contractual requirements and recovery not likely in a timely manner – corrective actions were ineffective	Red

The first category is Exceptional. An Exceptional rating means that "Performance meets contractual requirements and exceeds many to the Government's benefit. The contractual performance of the element or sub-element being assessed was accomplished with few minor problems for which corrective actions taken by the contractor were highly effective" (Gansler, 1997:9). This rating corresponds with Blue for the CPARS.

The second type of rating is Very Good. The definition of "Very Good" means, "Performance meets contractual requirements and exceeds some to the Government's benefit. The contractual performance of the element or sub-element being assessed was accomplished with some minor problems for which corrective actions taken by the contractor were effective" (Gansler, 1997:9). This corresponds to a Purple CPAR mark. With this new CPAR color rating, Purple, a Blue CPAR rating should be reserved for only truly outstanding performance (Hanson, 1998:9).

The middle classification is called Satisfactory. By receiving a Green rating, the contractor's "Performance meets contractual requirements. The contractual performance of the element or sub-element contains some minor problems for which corrective actions taken by the contractor appear or were satisfactory" (Gansler, 1997:9). A Satisfactory is equivalent to Green for the CPARS.

The fourth grade is Marginal. A Yellow rating states that, "Performance does not meet some contractual requirements. The contractual performance of the element or sub-element being assessed reflects a serious problem for which the contractor has not yet identified corrective actions. The contractor's proposed actions appear only marginally effective or were not fully implemented" (Gansler, 1997:9). Yellow in the CPAR format equates to a Marginal rating.

The final category is Unsatisfactory. An Unsatisfactory is warranted if, "Performance does not meet most contractual requirements and recovery is not likely in a timely manner. The contractual performance of the element or sub-element contains serious problems for which corrective actions taken by the contractor appear or were ineffective" (Gansler, 1997:9). A poor score of Unsatisfactory is depicted by a Red CPARS rating.

The relationship between PPI, CPARS, the assessment elements and measurable data is shown by Figure 4. Typically, source selections evaluate proposals through a factor assessment, which is a combination of cost, specific, and assessment criteria, proposal risks, and performance risks (Wright, 1997:17). Performance risks is based on the bidders past and present performance (Wright, 1997:19) which is PPI. Again, PPI has many inputs, one of which is the CPARS. The CPARS database contains the information based on thresholds and business sector as discussed earlier. Finally, the Cost Control rating, for example, must come from a documented source such as cost performance reports (CPR) (AFMC, 1997: 1). This example illustrates how an objective measurement must be used for the basis of PPI.

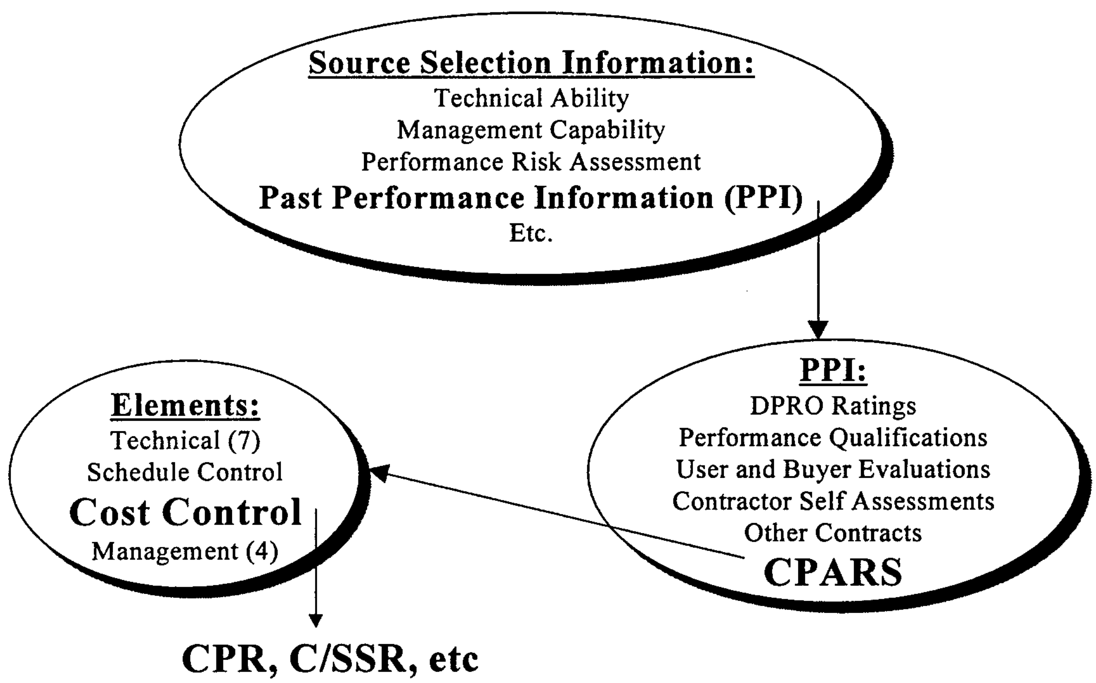

Figure 4. How Measurable Data Relates to PPI

Thus, for AF source selections, CPARS ratings supply the integral element of PPI that feeds into part of the source selection criteria. The Source Selection Authority then uses the criteria in aggregate form to select the best value vendor in accordance with the Source Selection Evaluation Plan (FAR, 1998:4). The award of new contracts then affects the corporation performance measurements discussed earlier. The cycle then starts again.

Summary

The government has attempted to use PPI during source selection evaluations several times and has determined that the costs of collecting and using PPI did not outweigh the benefits achieved during source selections. Nonetheless, current DoD policy is reinforcing this Industry practice of evaluating possible vendors' prior effort on similar contracts. The Little study has concluded that using past performance data for DoD source selections does make sense and is being used successfully, although in a constrained manner (Little, 1996:10). AFMC has designated the CPARS to be the PPI vehicle used for AFMC acquisitions. These acquisitions can be categorized into five sectors: Systems, Services, Operations Support, Information Technology, and Unique Business sectors. In order for the CPARS or any other PPI system to be effective, the Little report states that the data must be reliable (Little, 1996:89). Because PPI information is used to predict the best value vendor, the primary consideration for the AF using the CPARS concerns its reliability.

Propositions

This thesis uses the following propositions in an effort to determine the predictive reliability of the CPARS.

- Proposition: Cost and schedule variances are the primary determinants of the cost and schedule color ratings.
- Proposition: Although overlap exists between adjacent colors, mean cost and schedule variances are different for each color rating. A contractor who performs at a given level in terms of cost and schedule variance should receive a corresponding rating.
- Proposition: The reliability of the CPARS has changed over time.
- Proposition: Contractors who earn good ratings enjoy the highest profitability ratings. In other words, the best contractors make the most money. Also, DoD policy rewards good performance with higher profits.

In the following chapter the propositions listed above will be more clearly delineated in the form of hypotheses that can be tested using the presented methodology.

III. Methodology

Overview

This chapter describes the methodology used to answer the investigative questions presented in Chapter I. Recall the problem statement is to investigate the reliability of the CPARS. Answering Investigative Questions #1 and #2 through Hypotheses 1-6, the reliability of the CPARS can be determined. Together, the answers to the first two questions will help decide the system's overall reliability. The third and final Investigative Question will help judge if the perceived "best" contractors in terms of factor ratings are the most "profitable" in terms of Return on Investment (ROI) percentage and Return on Equity (ROE) percentage.

This study will rely on several statistical techniques. First, a correlation analysis using Spearman's rank correlation coefficient will be performed to determine the relationship between a categorical variable, usually color rating, and an objective measurement such as cost and schedule variances. The second test is a Tukey multiple comparison procedure. This procedure will determine if the means of the objective measurements are different for each of the categorical variables. Finally, a simple linear regression model will be implemented to answer two of the hypotheses regarding historical trends.

Data Collection

The bulk of the data has been collected from the ASC portion of the CPARS database. The CPARS database is owned and maintained by AFMC. Representatives at the product centers are responsible for entering the data provided by the SPOs from their respective centers. The CPAR System was established in 1988 by the Air Force (Sumpter, 1998:3). Data contained in the database includes current contracts as well as contracts that have been completed within the last three years. The CPARS database available for this study contains color ratings, cost, schedule, and technical performance data (performance not addressed in this thesis), and contract information from September 1988 through April 1998. Currently the CPARS database contains nearly 3,000 records. The data collected for this study includes all Cost-type contracts of the CPARS database that begin with F33657 (denotes ASC Contracts). Limiting the scope to ASC Cost-type contracts bounded the data used in this study to a maximum of 149 records. While reducing the sample significantly limits the results of this thesis, ASC data is an outstanding sample of major system's contracting due to size and complexity of acquisition programs.

Data was also collected from the ASC Cost/Schedule Data Center (ASC Cost Library). Pertinent data contained in the library includes Cost Performance Reports (CPR) and Cost/Schedule Status Reports (C/SSR) (ASC, 1996:ii). However, library data provided by System Program Offices (SPOs) was neither extensive nor consistent from contract to contract. Therefore, an attempt to obtain missing library data was made by contacting local SPOs, if the office still existed. Data was extracted from the cost reports because the current CPARS policy is to report *cumulative* cost and schedule variances and not *period* cost and schedule variances (AFMC, 1997:17), even though the color ratings should apply to the rating period only.

The final piece of the data was found on the Internet. Profitability information of the corporations was taken from the Morningstar web site, which refers to each company's Annual reports. ROI percentage and ROE percentage were chosen to represent financial measures of "fitness" since they help "define one set of necessary conditions for 'excellence'" (Chakravarthy, 1986:455). Information could not be located for each contractor from the Morningstar web site, which further limited the sample size when testing Investigative Question #3.

The data used for this research is contained in the Appendix: Data Tables. The data includes a number assigned to each CAGE Code and a letter assigned to each different contract number. The CAGE Codes and contract numbers are masked to protect the identity of contractors. The next two fields are the beginning and ending of the rating period. Cost and schedule data were taken from these months to establish period performances, which will be discussed later. Percent Complete is the final block before the rating information. This information was used to eliminate some of the points when conducting schedule tests. SV%, by definition, approaches zero as the contract approaches completion. It does not make sense then to evaluate the SV% of contracts that are at or near 100% complete. Finally, the color ratings and a cumulative variance reported in the CPARS database and a period variance calculated from CPRs and C/SSRs are listed for cost and schedule. The financial data obtained from the Internet was not included in the Appendix in order to avoid risking any corporate identification.

Statement of Hypotheses

The investigative questions and hypotheses tested are listed below.

Investigative Question #1

Is the CPARS reliable? Do performance measures positively correlate with performance ratings? The answer to this question must be determined by first answering two more specific questions. The first question compares the ratings at an aggregate level with objective measurements. The next part determines whether the average objective measurements are actually different across the color rating scale.

a. Do the ratings have a positive correlation with objective measurements? For the first two hypotheses, if the null hypothesis, H_0, cannot be rejected, then objective CPR or C/SSR data is not primarily determining the cost and schedule ratings. If the null is rejected, then objective measures such as cost and schedule variances may be indicators of a contractor's rating. A positive correlation would indicate that as cost and schedule variances improve, so does the color rating.

Hypothesis 1:

H_0: There is no correlation between the ratings and period CV%.
H_1: There exists a correlation between the ratings and period CV%.

Hypothesis 2:

H_0: There is no correlation between the ratings and period SV%.
H_1: There exists a correlation between the ratings and period SV%.

b. Are there differences in the mean objective measurements between each rating category? If the null hypothesis, H_0, cannot be rejected, then there is no statistical difference between average objective measure for each category. In other words, two contractors with a given cost or schedule variance can receive any color rating. If the null is rejected, then there is a statistically significant difference between the value of at least two color ratings.

Hypothesis 3:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean CV% for each color rating.
H_1: At least one mean is different.

Hypothesis 4:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean SV% for each color rating.
H_1: At least one mean is different.

Investigative Question #2:

Has the CPARS reliability changed over time? Hypotheses 5 and 6 will first determine a Spearman's Rank Correlation Coefficient for each period. A line will then fit to these coefficients using a simple, linear Least-Squares-Best-Fit model. The coefficient value of the independent variable (time) will provide information of the CPARS' reliability over time. If the coefficient is a positive (negative) number, then the reliability of the CPARS is improving (worsening) with respect to time. If the coefficient is approximately zero, then the reliability of the process has not changed over time. The p-value of the slope will determine if the change in the line, and thus the change in reliability of the CPARS, is significant.

Hypothesis 5:

H_0: The relationship between ratings and period CV% has not changed over time.
H_1: The relationship between ratings and period CV% has changed over time.

Hypothesis 6:

H_0: The relationship between ratings and period SV% has not changed over time.
H_1: The relationship between ratings and period SV% has changed over time.

Investigative Question #3

Do the past performance ratings correlate with measures of profitability? Are the perceived "best" contractors actually the most "profitable"? This question is intended specifically to address an Industry concern that "Excellent" performance has a cost and thus decreases profits. As in Investigative Question #1, this question must be broken into two different questions. The first question is concerned with correlation between the color ratings and objective measures and the period objective measurement (in this case, corporate ROE% and ROI%). The second question decides whether mean contractor profitability is different between color ratings. Finally, the findings from the reliability questions must be considered before answering these hypotheses. Clearly, if the cost or schedule ratings are found to be unreliable, then it makes no sense to find correlations with invalid systems. Therefore, these questions depend upon the results of the first two

Investigative Questions. If the CPARS are found to be unreliable for either the cost or schedule ratings, then finding their correlation with profitability measures is a moot exercise.

a. Do the ratings have a positive correlation with profitability measurements? The next hypotheses, 7 through 10, investigate the relationship between CPAR ratings and profitability. Failing to reject the null hypothesis indicates that there is no correlation between the CPARS color ratings and profitability measures. If this correlation is positive, then higher ratings coincide with higher profits. If this correlation is negative, then higher ratings correspond with lower profits. For hypotheses 11 through 14, a failure to reject the null hypotheses implies objective CPR or C/SSR contract data is not correlated with corporate profitability. If the null is rejected, then objective CPR or C/SSR contract data is correlated with corporate profitability

Hypothesis 7:

H_0: There is no correlation between the cost rating and ROE%.
H_1: There exists a correlation between the cost rating and ROE%.

Hypothesis 8:

H_0: There is no correlation between the cost rating and ROI%.
H_1: There exists a correlation between the cost rating and ROI%.

Hypothesis 9:

H_0: There is no correlation between the schedule rating and ROE%.
H_1: There exists a correlation between the schedule rating and ROE%.

Hypothesis 10:

H_0: There is no correlation between the schedule rating and ROI%.
H_1: There exists a correlation between the schedule rating and ROI%.

Hypothesis 11:

H_0: There is no correlation between the period CV% and ROE%.
H_1: There exists a correlation between the period CV% and ROE%.

Hypothesis 12:

H_0: There is no correlation between the period CV% and ROI%.
H_1: There exists a correlation between the period CV% and ROI%.

Hypothesis 13:

H_0: There is no correlation between the period SV% and ROE%.
H_1: There exists a correlation between the period SV% and ROE%.

Hypothesis 14:

H_0: There is no correlation between the period SV% and ROI%.
H_1: There exists a correlation between the period SV% and ROI%.

b. Are there differences in the mean profitability measurements between each rating category? For hypotheses 15 through 18, if the null hypothesis cannot be rejected, then there is no statistical difference between profitability measures for each rating category. If the null is rejected, then there is a statistical difference between the average profitability value between at least one color rating.

Hypothesis 15:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean ROE% for each cost rating.
H_1: At least one mean is different.

Hypothesis 16:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean ROI% for each cost rating.
H_1: At least one mean is different.

Hypothesis 17:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean ROE% for each schedule rating.
H_1: At least one mean is different.

Hypothesis 18:

H_0: $\mu_{Blue} = \mu_{Green} = \mu_{Yellow} = \mu_{Red}$; where μ_I = mean ROI% for each schedule rating.
H_1: At least one mean is different.

Method of Analysis

Each of the hypotheses generated from the three investigative questions will be primarily tested using either a test for correlation or a test for differences between population means. A regression analysis will be performed for Hypotheses 5 and 6.

Correlation Test

"Correlation models are employed to study the nature of the relations between the variables; they also may be used for making inferences about any one of the variables on the basis of the others" (Neter and others, 1996: 631). A correlational relationship is a relationship in which there is no direct control of the variables possessed by the items being studied (Kachigan, 1991:118). Five basic types of correlational relationships exist: linear positive, linear negative, nonlinear (curvilinear), cyclical, or no relation (independent) (Kachigan, 1991:119-120). This study will assume that all relationships analyzed will be linear in nature. Figure 5 illustrates four of these relational types.

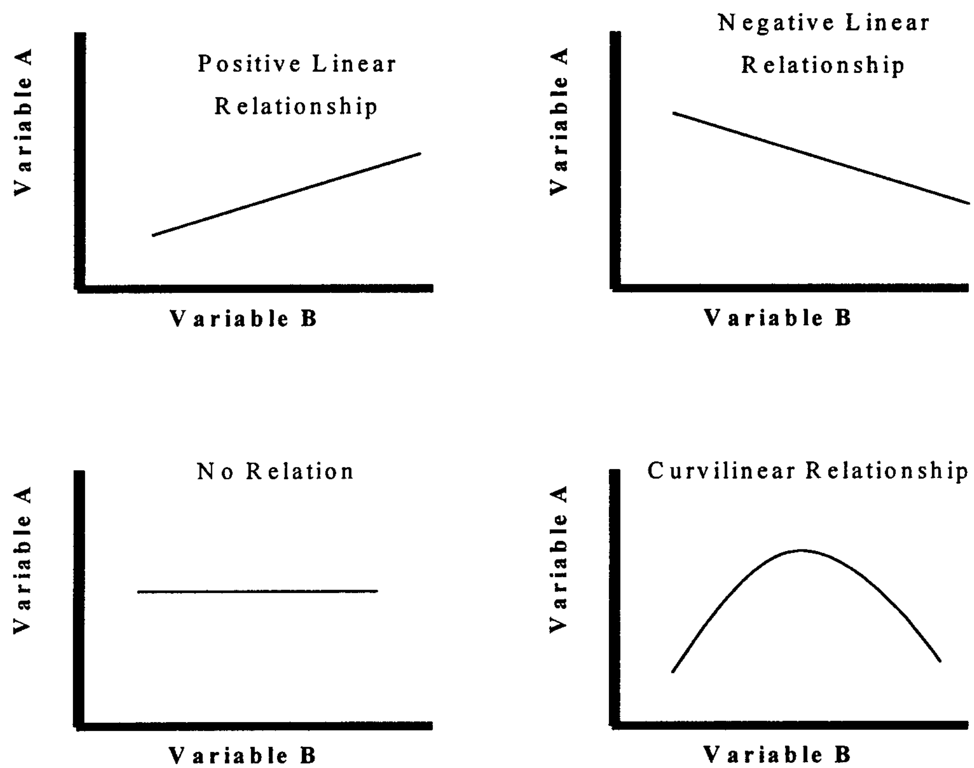

Figure 5. Correlational Relationship Types (Kachigan, 1991:121)

The correlation coefficient of a given sample is described by the letter r and estimates the population correlation coefficient, ρ (rho). Also, r can assume any value in the range –1.00 to 1.00 (Kachigan, 1991:126). A value of 1.00 means a perfect positive correlation exists between the two variables. Likewise, an r-value of –1.00 is defined as a perfect negative correlation. A value of 0 implies that there is no linear relationship between the two variables being studied.

The Pearson's product moment correlation coefficient is a estimator of the population correlation coefficient, ρ (rho) (Neter and others, 1996: 641). The Pearson coefficient is used when the joint distribution of the two random variables is a bivariate normal distribution. However, no known transformations exist to transform the ordinal data used in this thesis to normal, continuous data. "When no appropriate transformations can be found, a nonparametric *rank correlation* procedure is useful for making inferences about the association between Y_1 and Y_2. The *Spearman rank correlation coefficient* is widely used for this purpose" (Neter and others, 1996: 651). Thus, for this effort, a Spearman rank correlation coefficient will be used as the primary evaluator in all correlation tests.

To find the Spearman coefficient, the data in both categories must be assigned a rank. The ranks are labeled R_{i1} and R_{i2} for the two categories investigated. In the event of ties, each of the tied values is given the average of the ranks of the tied values involved. The Spearman rank correlation coefficient, r_S, is defined as:

$$r_S = \frac{\sum (R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\left[\sum (R_{i1} - \bar{R}_1)^2 * \sum (R_{i1} - \bar{R}_1)^2 \right]^{\frac{1}{2}}}$$

The Spearman rank correlation coefficient can be used to test the hypothesis:

H_0: There is no correlation between Y_1 and Y_2.
H_a: There exists a correlation between Y_1 and Y_2 (Neter and others, 1996: 651).

The probability distribution of the Spearman coefficient is "based on the condition that, for any ranking of Y_1, all rankings of Y_2 are equally likely when there is no association between Y_1 and Y_2" (Neter and others, 1996: 651). When the sample size, n, is greater than ten, the above hypothesis test can be conducted using the test statistic:

$$t^* = \frac{r_S \sqrt{n - 2}}{\sqrt{1 - r_S^2}}$$

This statistic is based on the t-distribution with n-2 degrees of freedom (Neter and others, 1996: 652). If t* is greater than the t-value from the t-distribution, then reject the null hypothesis.

When using correlations, it is important to note that correlation does not imply causality. For example, a city's telephone booths and population usually are strongly correlated, but adding or removing phone booths does not cause the population to increase or decrease.

Tukey Multiple Comparison Test

Analysis of Variance, or ANOVA, is a collection of techniques useful, "for identifying and measuring the various sources of variation within a collection of data" (Kachigan, 1991:195). The test statistic used with ANOVA tests is compared with the F-distribution. More specifically, the F test statistic, is the ratio of the mean square error for treatments (MSTr) to the mean square error (MSE), or common variance, of the entire sample. If the means of each category are close tot the overall mean, then the ratio will be reasonably close to one. However, if the treatment means begin to deviate from the overall mean, then the ratio will become larger (Devore, 1991:374-376).

Often, when a single-factor ANOVA fails an F-test, the analysis is terminated. However, when the null hypothesis is rejected, knowing which of the means are different becomes useful information. One method for conducting this further research is called a Tukey multiple comparison procedure (Devore, 1991:381).

"If an ANOVA experiment involves comparison of four treatments, then Tukey's procedure obtains simultaneously six different intervals" (Devore, 1991:384). The alpha error rate, or Type I error rate, no longer concerns one particular interval, but instead refers to the experiment as a whole. The error rate for each of the intervals must be lower (wider confidence intervals). Thus the alpha error rate is called an experimentwise error rate (Devore, 1991:384). This procedure enables the researcher to examine "all pairwise group differences on a variable with experimentwise error rate held in check" (Stevens, 1992:203).

Minitab will be used to perform the Tukey's multiple comparison procedure. The software will simultaneously determine confidence intervals for the six different groups so that the experimentwise Type 1 error is 5%. "If the confidence interval includes 0, we conclude the population means are not significantly different" (Stevens, 1992:203). This is due to the fact that if the interval includes zero, then zero is a likely solution to the equation, $\mu_i - \mu_j = X$, which would be equivalent to $\mu_i = \mu_j$ (Stevens, 1992:203). Thus, two items will be of interest from the Minitab output: the p-value, which indicates the smallest α-value for which the null hypothesis can be rejected, and the Tukey intervals that identify which interval means are different.

Regression Test

Regression will be employed to discover whether the correlation between color ratings and objective measures, cost and schedule variance, have changed over time. A higher correlation of color ratings to specific performance measures implies that the system is more reliable. If reliability of the CPARS is improving over time, then the correlations will be different *and* increasing over time. Thus, the Spearman rank correlation coefficient will be regressed against periods of time.

"The simplest deterministic mathematical relationship between two variables x and y is a linear relationship $y = \beta_0 + \beta_1 x$" (Devore, 1991:454). Regression analysis can be used to describe, control, or predict the relationship between two or more variables (Neter and others, 1996:9). The parameters β_0 and β_1 are called regression coefficients and are estimated by b_0 and b_1. The population parameter, β_1, is the slope of the line (Neter and others, 1996:12,20). Thus, if β_1 is not equal to zero, then the correlations between color ratings and objective performance measurements are changing. The sign of the slope determines whether the change in correlation over time is positive or negative.

The simple linear regression model is often used to determine whether or not there is a linear association between two variables. The two alternatives are:

H_0: $\beta_1 = 0$.
H_a: $\beta_1 \neq 0$ (Neter and others, 1996: 51).

An explicit test of the alternatives is based on the test statistic:

$$t^* = \frac{b_1}{s\{b_1\}}$$

The decision rule is similar to the correlation test. If $|t^*|$ is greater than the t-value obtained from the t-distribution, then reject the null hypothesis (Neter and others, 1996: 51).

Data Preparation

For each test, notable "data preparation" will be necessary. The bulk of the preparation of the data is basically the same for each test. First, eliminate extraneous fields. Second, eliminate any records that did not contain data. Outliers in the data will be eliminated if and only if they are extreme cases or caused by policy changes such as re-baselining.

Period data was obtained by researching CPR and C/SSR information. The cumulative Budgeted Cost of Work Scheduled (BCWS), Budgeted Cost of Work Performed (BCWP), and Actual Cost of Work Performed (ACWP) were taken from contract cost information from the beginning and ending of each CPARS reporting period. The period ending numbers were subtracted from the period beginning numbers to obtain a period BCWS, period BCWP, and period ACWP. These values were then entered into the standard cost and schedule variance percentage equations that provided the period cost and schedule variance percentages.

For Spearman correlation tests, the data will be assigned a numerical value based on the rank of the observation. Similar color ratings will be treated as ties and will be assigned the average of all the ranks for that rating. For the Tukey multiple comparison tests, the appropriate data columns will be copied from Microsoft Excel to a Minitab worksheet and evaluated in Minitab. For the regression tests, the data will be divided in an annual or biannual fashion while maintaining a minimum group size of 10.

Robustness

To add robustness to this study, all hypotheses with correlation will be analyzed using the Pearson's correlation briefly discussed in the Correlation Tests section. The Spearman and Pearson correlation coefficients each demand different properties of the data. Because of the differences, the strength of the conclusions is increased if both correlations are "close" to one another. Only the Spearman correlation will be calculated for the regression portion of the study.

For this study, p-values will be reported for each test. The p-value is defined as the smallest α for which the null hypothesis can be rejected. For each hypothesis test, a p-value less than 0.05 will indicate strong support for the rejection of the null hypothesis, H_0. These situations will be referred to as "the p-value strongly supports the rejection of the null hypothesis," or other comparable wording. Likewise, a p-value between 0.10 and 0.05 will suggest a moderate support for the rejection of the null hypothesis, H_0. Its phrasing will be worded similarly. Also, tests will be reevaluated after identifying and removing extreme cases, or outliers.

Limitations

As with any non-experimental study, sample size is a significant limitation. Using only ASC, Cost-type contracts significantly limits the database. Larger sample sizes increase the strength of any test and this weakness is the most glaring during the Regression tests. The small sample size also detracts from having a balanced design, which is most notable for the Red category. Another limitation is the fact that period CV% and SV% were not taken from the exact beginning day and exact ending day of the period. Instead, there were often a week or two, and occasionally several months (the closest report that could be found) between the financial report used and the CPARS period dates. Next, the percentage completion for contracts to evaluate schedule ratings was an arbitrary selection. Only contracts less than or equal to 80% complete were included in this study in an effort to preserve sample size while eliminating nearly completed contracts (again, because that SV% approaches zero near completion). Another limitation is presented with the ROE and ROI percentages. These percentages are corporate level percentages, not contract or CAGE Code specific.

Summary

This study will rely primarily on correlation analyses, multiple comparison procedures, and regression techniques to answer the hypotheses. By answering the outlined hypotheses, the reliability of the CPARS process can be determined.

The data contained in the CPARS database provides the researcher the ability to compare performance color ratings against cumulative cost and schedule variances as well as other factors. The ASC Cost Library and SPO provided cost and schedule information augment the CPARS database. Together, these databases allow the investigation of the reliability of the CPARS process and its relationship with corporate profitability.

IV. Findings

Overview

This chapter presents the results of the statistical tests conducted to answer the three investigative questions. The results are presented in order of the investigative questions answered.

Investigative Question #1

Hypotheses 1 and 2

Both the Spearman and Pearson correlations between the period CV% and cost color ratings were found to be nearly equal. Their p-value scores were also within the same range. Both the Spearman p-value and the Pearson p-value moderately support the rejection of the null hypothesis. Thus, there is moderate support that aggregate period CV% and cost color ratings are slightly correlated. For period SV%, the correlations with schedule color ratings were again found to be similar. Their p-value scores were also somewhat comparable even though neither supports the rejection of the null hypothesis. Therefore, there is no support that aggregate period SV% and schedule color ratings are correlated. Table 5 shows the results of the first two hypothesis tests.

Table 5. Correlations for Hypotheses 1 and 2

	Spearman's Correlation	**p-value**	**Pearson's Correlation**	**p-value**
H1	0.209	0.061	0.224	0.096
H2	-0.150	0.178	-0.196	0.225

Hypothesis 3

The ANOVA p-value of 0.170 does not support the rejection of the null hypothesis. The implication then is that the cost color ratings do not discriminate between contractors' performance using period CV%. Also, each of the intervals in Figure 6 contains zero, which supports the implication that there is no difference between the mean CV% for each of the color ratings. Figure 7 shows that there are no extreme departures from normality.

```
Analysis of Variance for CV%
Source      DF        SS         MS         F        P
Cost         3      1783        594      1.74    0.170
Error       52     17730        341
Total       55     19513
                                  Individual 95% CIs For Mean
                                  Based on Pooled StDev
Level        N      Mean     StDev  --------+---------+---------+--------
Blue        17      6.73     13.07                    (-------*------)
Green       23     -6.15     18.92        (------*-----)
Yellow      12     -4.84     16.17       (--------*--------)
Red          4     -2.27     37.63   (--------------*---------------)
                                    --------+---------+---------+--------
Pooled StDev =      18.47                 -12         0        12

Intervals for (column level mean) - (row level mean)

                 Blue        Green        Red
  Green         -2.78
                28.54

  Red          -18.21       -30.41
                36.21        22.64

  Yellow        -6.89       -18.74       -25.70
                30.03        16.13        30.84
```

Figure 6. Hypothesis 3 Tukey Intervals

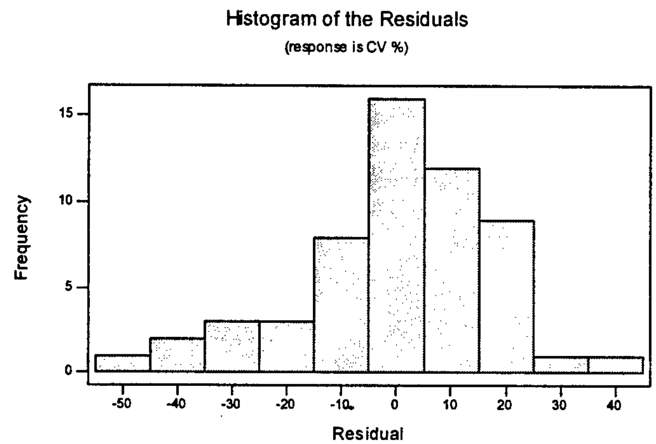

Figure 7. Hypothesis 3 Residual Histogram

Hypothesis 4

The ANOVA p-value of 0.131 does not support that the schedule color ratings discriminate between contractors' performance using period SV%. The intervals in Figure 8 all contain zero. However, the Residual Histogram in Figure 9 indicates that there may be an outlier. The Normal Probability Plot in Figure 10 supports that one point is an extreme observation. The point was removed and the model was rerun.

For the re-evaluation without the outlier, the ANOVA p-value of 0.086 now moderately suggests the rejection of the null hypothesis. However, all of the intervals still contain zero. The negative correlation found by both the Spearman and Pearson correlations is illustrated in Figure 11 by the fact that the average SV% of the Green rating is less than the average SV% of the Yellow rating. The cost color ratings do not discriminate between contractors' performance using period CV%. Figure 12 shows that there are no extreme departures from normality.

```
Analysis of Variance for SV%
Source      DF        SS        MS        F        P
Schedule     3       984       328     2.00    0.131
Error       36      5904       164
Total       39      6888
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level        N      Mean     StDev --------+---------+---------+--------
Blue         4      0.66      5.75           (--------*--------)
Green       26     -7.94     13.77          (---*--)
Yellow       9      3.32     11.53                (-----*-----)
Red          1      0.00      0.00  (----------------*----------------)
                                   --------+---------+---------+--------
Pooled StDev =     12.81                 -15          0         15

Intervals for (column level mean) - (row level mean)

               Blue        Green        Red
  Green        -9.92
               27.13

  Red         -37.91      -43.10
               39.23       27.21

  Yellow      -23.39      -24.60      -39.68
               18.07        2.08       33.05
```

Figure 8. Hypothesis 4 Tukey Intervals

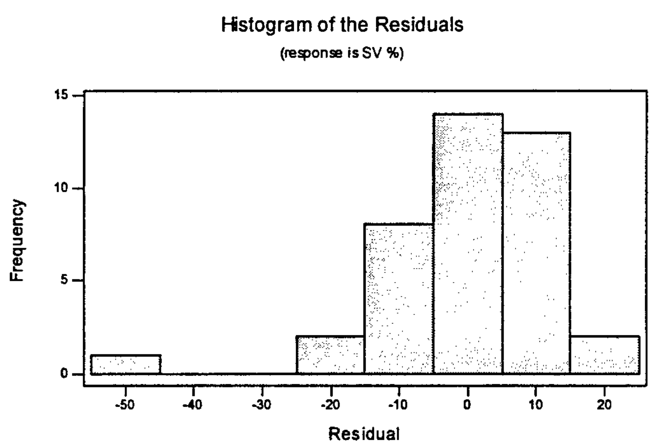

Figure 9. Hypothesis 4 Residual Histogram

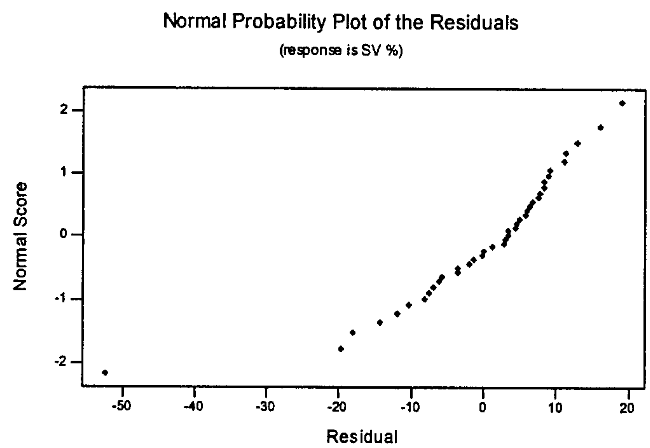

Figure 10. Hypothesis 4 Normal Probability Plot

```
Analysis of Variance for SV%
Source     DF        SS         MS         F        P
Schedule    3     624.7      208.2      2.39    0.086
Error      35    3053.9       87.3
Total      38    3678.6
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  ------+---------+---------+---------+
Blue        4     0.660     5.750            (-------*------)
Green      25    -5.851     8.879           (--*--)
Yellow      9     3.318    11.526                 (-----*----)
Red         1     0.000     0.000  (---------------*---------------)
                                   ------+---------+---------+---------+
Pooled StDev =    9.341                 -12         0        12        24

Intervals for (column level mean) - (row level mean)

              Blue        Green        Red
  Green      -7.04
             20.06

  Red       -27.48       -31.51
             28.80        19.81

  Yellow    -17.78       -18.95       -29.84
             12.46         0.61        23.21
```

Figure 11. Hypothesis 4 Tukey Intervals (Outlier Removed)

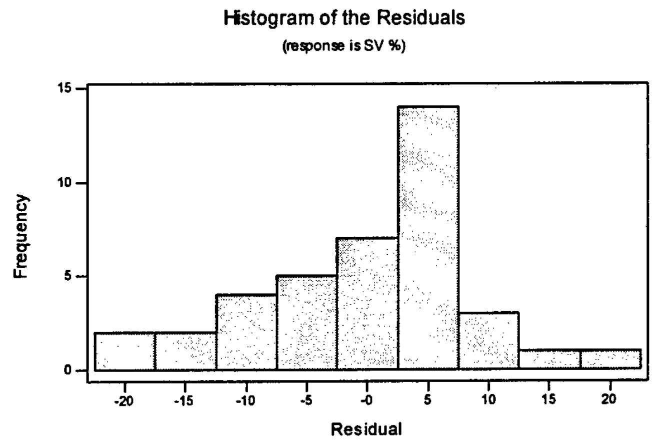

Figure 12. Hypothesis 4 Residual Histogram (Outlier Removed)

Investigative Question #1 (Using Cumulative CV% and SV%)

Because of the surprisingly poor results obtained for Investigative Question #1, the reliability of the CPARS process will be reassessed using cumulative CV% and SV% that is reported with the CPAR period evaluations.

Hypotheses 1* and 2* (Using Cumulative CV% and SV%)

Both the Spearman and Pearson correlations between the cumulative CV% and cost color ratings were similar. Also, both of the p-values strongly support the rejection of the null hypothesis. Thus, there is strong support that there is a moderate correlation between cumulative CV% and cost color ratings. In fact, these correlations were the highest aggregate correlations encountered during this study. The correlations and p-values for cumulative CV% were considerably better than for period CV%. The implication of this result will be discussed more fully in the next chapter.

For cumulative SV%, both the Spearman and Pearson correlations were not as similar, yet their results were the same. Neither of their p-values supports the rejection of the null hypothesis. Thus, there is no support that there is a correlation between cumulative SV% and schedule color ratings. Table 6 shows the results of these two tests.

Table 6. Correlations for Hypotheses 1* and 2* (Cumulative Measures)

	Spearman's Correlation	**p-value**	**Pearson's Correlation**	**p-value**
H1*	0.524	<0.00001	0.447	<0.001
H2*	0.141	0.106	0.034	0.770

Hypothesis 3* (Using Cumulative CV% and SV%)

Figure 13 shows that the ANOVA p-value strongly supports the rejection of the null hypothesis and several intervals do not contain zero. However, Figures 14 and 15 indicate that an outlier may be present in the data. This outlier was identified and removed.

After removing the outlier, the ANOVA p-value still strongly supports the rejection of the null hypothesis. Adjacent color ratings did overlap, but non-adjacent color ratings did not overlap. Figure 16 shows the Tukey comparison output for Hypotheses 3* after the removal of the outlier. Figure 17 displays what initially looks like potential outliers; however, since the two points are symmetric and the results are significant, they are assumed to be in the tails of the normal residual distribution and were not removed from the analysis. The analysis confirms that there are differences between the population means of cost color ratings when using cumulative CV%.

Furthermore, these means are ordered in an appropriate descending manner, from Blue to Red.

```
Analysis of Variance for CV%
Source      DF       SS        MS        F        P
Cost         3    13335      4445    12.79    0.000
Error      116    40310       347
Total      119    53645
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean     StDev  --------+---------+---------+--------
Blue       37      7.19      8.84                            (--*---)
Green      40     -0.35     12.64                       (---*--)
Yellow     33     -8.14     12.49                  (---*---)
Red        10    -31.89     54.04   (------*------)
                                   --------+---------+---------+--------
Pooled StDev =    18.64                   -32       -16         0

Intervals for (column level mean) - (row level mean)

            Blue         Green        Red
  Green    -3.55
           18.64

  Red      21.74         14.34
           56.41         48.73

  Yellow    3.69         -3.65      -41.30
           26.98         19.23       -6.19
```

Figure 13. Hypothesis 3* Tukey Intervals (Cumulative CV%)

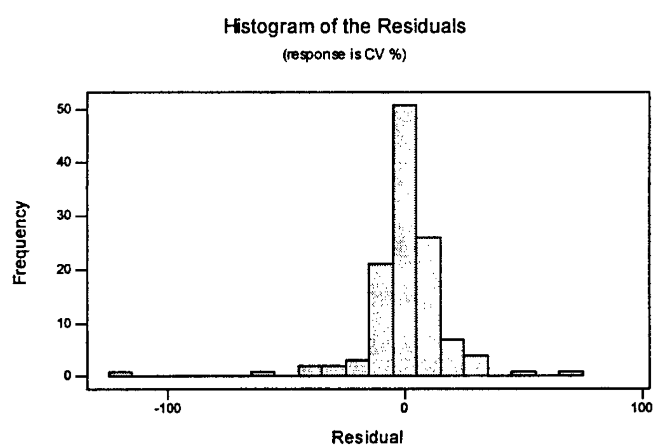

Figure 14. Hypothesis 3* Residual Histogram (Cumulative CV%)

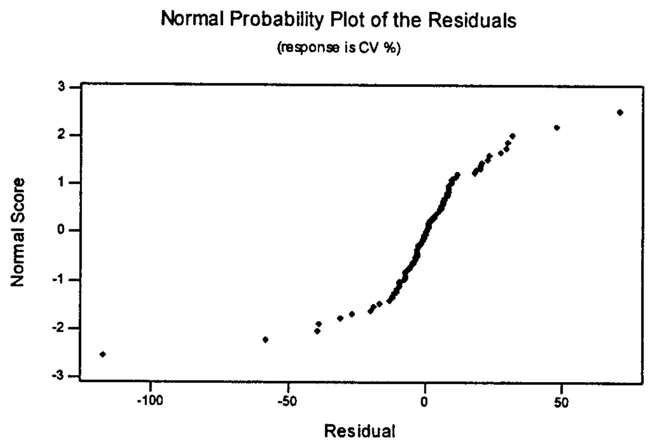

Figure 15. Hypothesis 3* Normal Probability Plot (Cumulative CV%)

```
Analysis of Variance for CV%
Source      DF        SS         MS        F        P
Cost         3      7002       2334    10.76    0.000
Error      115     24940        217
Total      118     31941
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level        N      Mean      StDev ----+---------+---------+---------+--
Blue        37      7.19       8.84                          (---*---)
Green       40     -0.35      12.64                    (---*---)
Yellow      33     -8.14      12.49              (---*---)
Red          9    -18.82      36.93 (-------*-------)
                                    ----+---------+---------+---------+--
Pooled StDev =     14.73               -24       -12         0        12

Intervals for (column level mean) - (row level mean)

             Blue        Green        Red
  Green     -1.22
            16.31

  Red       11.73         4.29
            40.29        32.64

  Yellow     6.13        -1.25      -25.13
            24.53        16.82        3.77
```

Figure 16. Hypothesis 3* Tukey Intervals (Cumulative CV% & Outlier Removed)

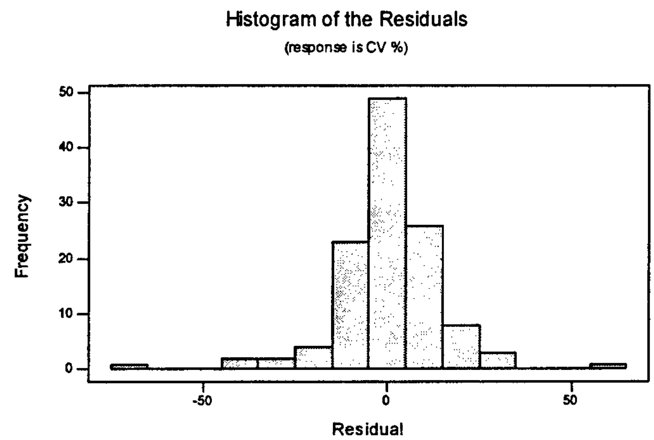

Figure 17. Hypothesis 3* Residual Histogram (Cumulative CV% & Outlier Removed)

Hypothesis 4*

The ANOVA test shows that mean cumulative SV% for each color rating is not significantly different. Figure 18 shows that the ANOVA p-value does not support the rejection of the null hypothesis and that all Tukey intervals contain zero. Also, Figure 19 indicates that outliers may be present in the data. Because of the extremely poor results of this model and the disbelief that the model would improve materially, the model was not re-evaluated without the potential outliers.

```
Analysis of Variance for SV%
Source     DF        SS         MS        F        P
Schedule    3       140         47     0.20    0.893
Error      74     16897        228
Total      77     17038
                                     Individual 95% CIs For Mean
                                     Based on Pooled StDev
Level      N      Mean      StDev  --------+---------+---------+--------
Blue      12     -5.12      10.75               (---*----)
Green     49     -2.98      15.86                  (--*-)
Yellow    16     -6.14      15.38               (---*---)
Red        1     -5.10       0.00  (--------------*---------------)
                                   --------+---------+---------+--------
Pooled StDev =     15.11                 -20         0        20

Intervals for (column level mean) - (row level mean)

             Blue         Green        Red
  Green     -14.94
             10.67

  Red       -41.39       -38.04
             41.35        42.27

  Yellow    -14.16        -8.29       -39.93
             16.20        14.60        42.01
```

Figure 18. Hypothesis 4* Tukey Intervals (Cumulative SV%)

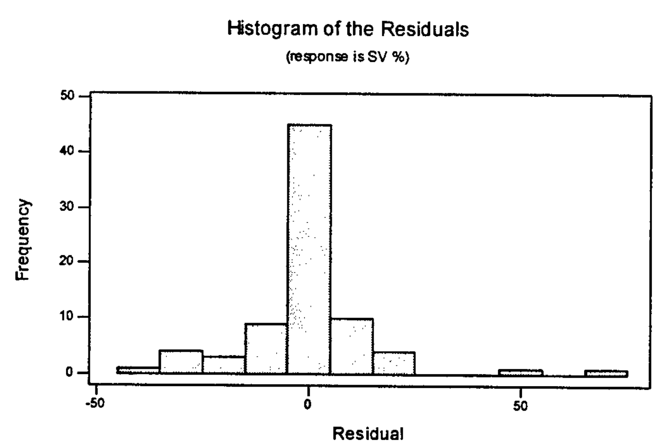

Figure 19. Hypothesis 4* Residual Histogram (Cumulative SV%)

Investigative Question #2

How has the reliability of CPARS changed over time? The purpose of the next two hypotheses was to statistically determine whether the reliability of the CPARS color ratings and their respective objective performance measures have changed over time.

Hypothesis 5

This hypothesis tested whether the reliability of the period CV% and cost color ratings has changed over time. The data was broken into four groups and a Spearman's correlation was calculated for each period. The period sizes, correlation values, and their respective p-values can be seen in Table 7.

Table 7. Correlations for Hypothesis 5

Period	N	Spearman's Correlation	p-value
Apr 92-Oct 92	12	0.183	.285
Feb 93–Oct 93	16	0.245	.180
Jul 94–Oct 95	15	0.378	.082
Jul 96–Sep 97	13	-0.061	.422

These correlation values were then plotted and regressed against time to determine whether the slope of the line had significantly changed. The p-value for the slope coefficient of the LSBF model for period CV% was 0.580. This value does not support the rejection of the null hypothesis. Therefore, the reliability of the CPARS color ratings against objective contract performance measures, period CV%, has not changed over time. The actual linear plot of the correlations is shown in Figure 20. The bold straight line is the trendline estimated by the Least Squares Best Fit (LSBF) linear model.

Figure 20 shows that the first three points appear to be increasing in a linear fashion. However, when the last point was removed, the p-value of the slope was still insignificant at $\rho = 0.1323$. Thus, the increase of the line considering only the first three points is not significantly different than zero.

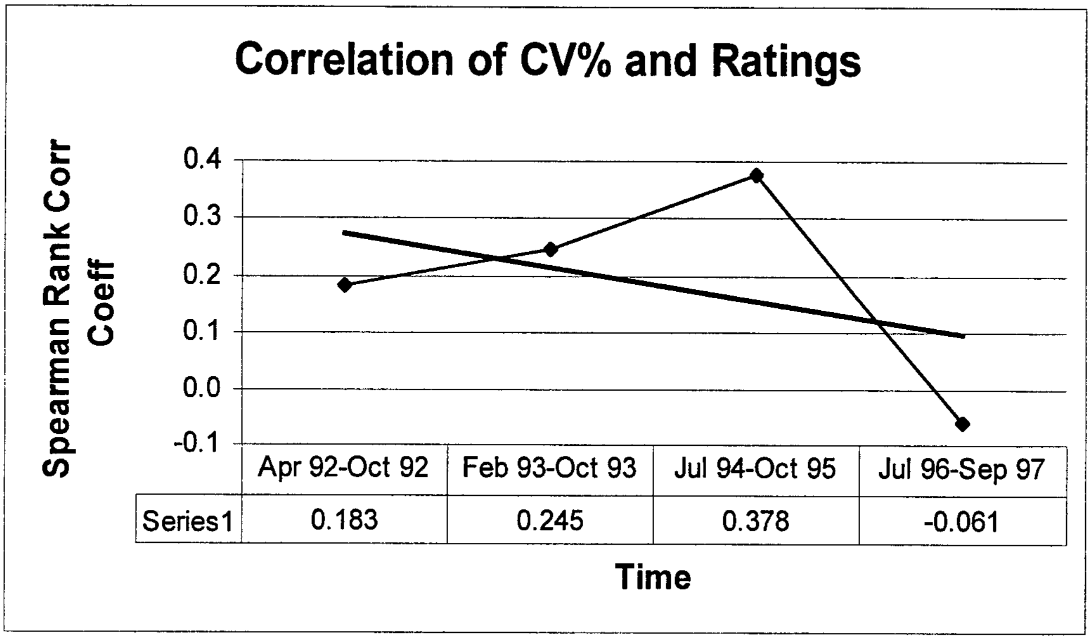

Figure 20. Period CV% and Ratings Correlation Trend

Hypothesis 6

This hypothesis tested whether the reliability of the period SV% and schedule color ratings has changed over time. The data was broken into three groups for this test. The calculated correlation values, which can be seen in Table 8, were then plotted and regressed against time to determine whether the slope of the line significantly changed.

Table 8. Correlations for Hypothesis 6

Period	N	Spearman's Correlation	p-value
Feb 90-Oct 92	11	-0.268	0.080
Mar 93–Jun 94	12	0.128	0.478
Sep 94-Jan 97	17	0.349	0.085

The p-value for the slope coefficient of the LSBF model was 0.030. This value strongly supports the rejection of the null hypothesis. Moreover, the positive value of the slope coefficient, $b_1 = 0.399$, indicates that the trend is positive. Therefore, the regression analysis provides strong support that the reliability of the CPARS schedule color ratings against objective contract performance measures, period SV%, has improved over time. Another interesting note is that the last period shows moderate support for a slight-to-moderate correlation given the same analysis criteria used in Hypothesis 2. The actual linear plot of the correlations is shown in Figure 21. The bold straight line is the trendline estimated by the Least Squares Best Fit (LSBF) linear model.

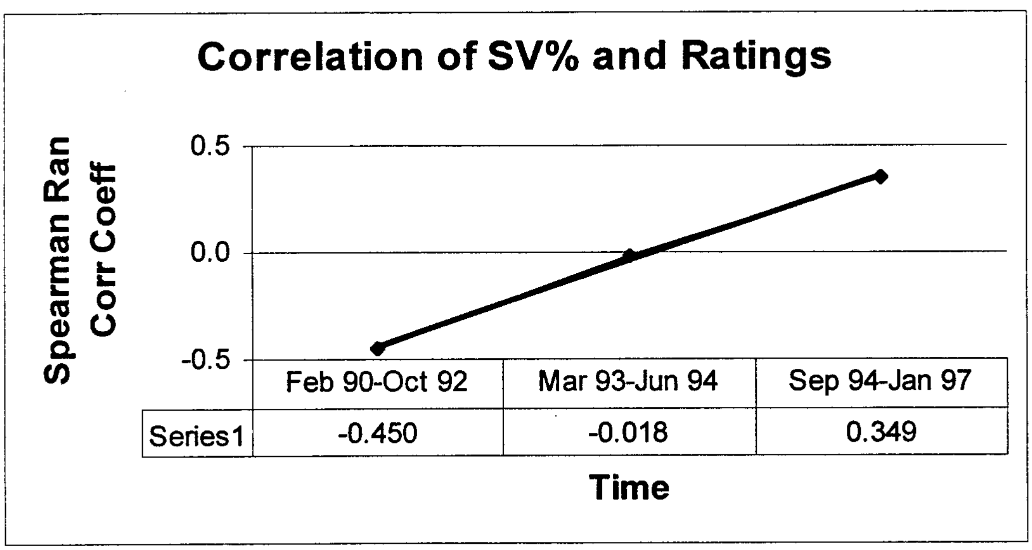

Figure 21. Period SV% and Ratings Correlation Trend

Investigative Question #2 (Using Cumulative CV% and SV%)

The next two hypothesis tests are exactly like Hypotheses 5 and 6, except the objective cost measure used is *cumulative* CV% and SV% instead of period CV% and SV%. The objective of these two hypotheses is to determine if the relationship between cumulative objective measures and color ratings have changed over time.

Hypothesis 5* (Using Cumulative CV% and SV%)

For this hypothesis, the data was broken into seven periods. A Spearman correlation was calculated for each period. The period sizes, correlation values, and their respective p-values are shown in Table 9.

Table 9. Correlations for Hypothesis 5*

Period	N	Spearman's Correlation	p-value
Sep 88-Oct 91	12	0.709	0.003
Apr 92-Oct 92	10	0.782	0.002
Feb 93-Oct 93	15	0.659	0.002
Jan 94-Sep 94	21	0.269	0.108
Oct 94-Oct 95	26	0.757	<0.00001
Jan 96-Sep 96	24	0.224	0.136
Oct 96-Jun 97	12	0.101	0.366

Regression of the correlation values against time was then performed to determine if the slope of the LSBF line had significantly changed. The p-value for the slope coefficient of the LSBF model for cumulative CV% was 0.049. This value strongly supports the rejection of the null hypothesis. Therefore, the reliability of the CPARS color ratings against objective contract performance measures, cumulative CV%, has changed over time. In addition, the negative value of the slope coefficient, $b_1 = -0.101$ indicates that the trend is negative; the reliability of the CPARS using cumulative CV% is weakening over time. The causes of this phenomenon will be discussed in the next chapter. The plot of the correlations against time is shown in Figure 22. The straight line is the trendline estimated by the regression equation.

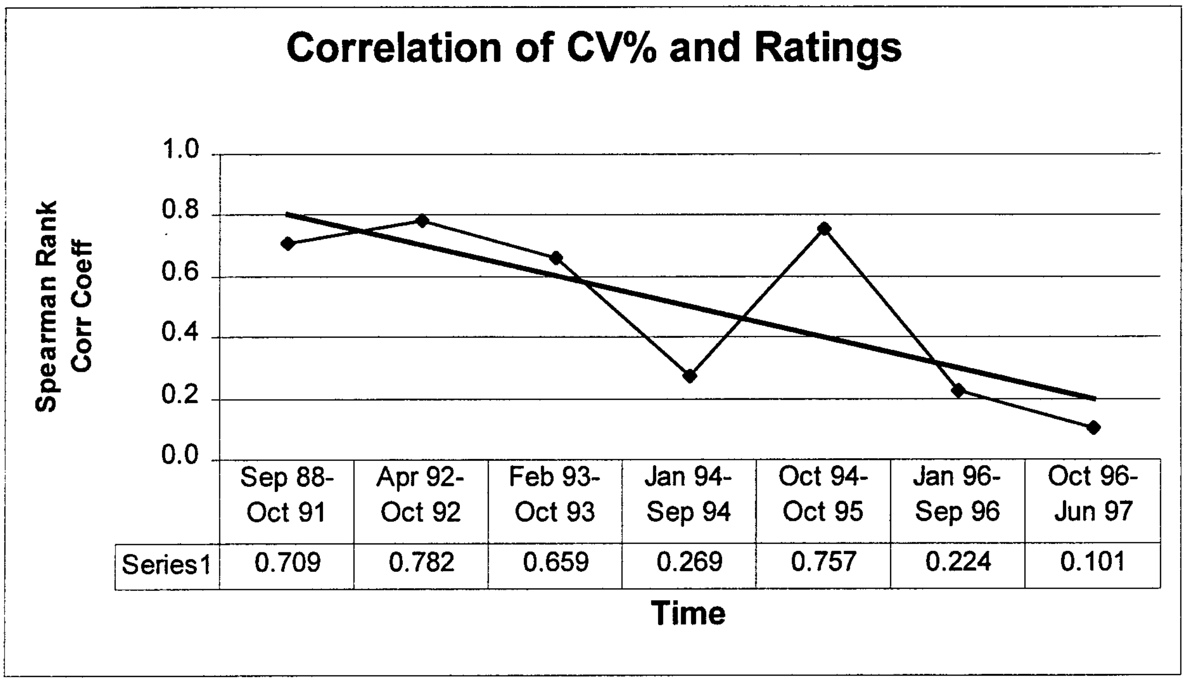

Figure 22. Cumulative CV% and Ratings Correlation Trend

Hypothesis 6* (Using Cumulative CV% and SV%)

The data was broken into six periods for this hypothesis. A Spearman correlation was calculated for each period. The period sizes, correlation values, and their respective p-values are shown in Table 10.

Regression of the correlation values against time was then performed to determine if the slope of the LSBF line had significantly changed. The p-value for the slope coefficient of the LSBF model for cumulative SV% was 0.303. This value does not support the rejection of the null hypothesis. Therefore, the reliability of the CPARS schedule ratings against objective contract performance measures, cumulative SV%, has not significantly changed over time. The plot of the correlations against time is shown in Figure 23. The straight line is the trendline estimated by the regression equation.

Table 10. Correlations for Hypothesis 6*

Period	N	Spearman's Correlation	p-value
Feb 90-Oct 91	10	0.340	0.143
Apr 92-Oct 92	11	0.140	0.325
Feb 93-Oct 93	13	-0.114	0.344
Mar 94-Oct 94	12	0.382	0.091
Dec 94-Oct 95	14	0.081	0.383
Jan 96-Feb 97	18	-0.143	0.275

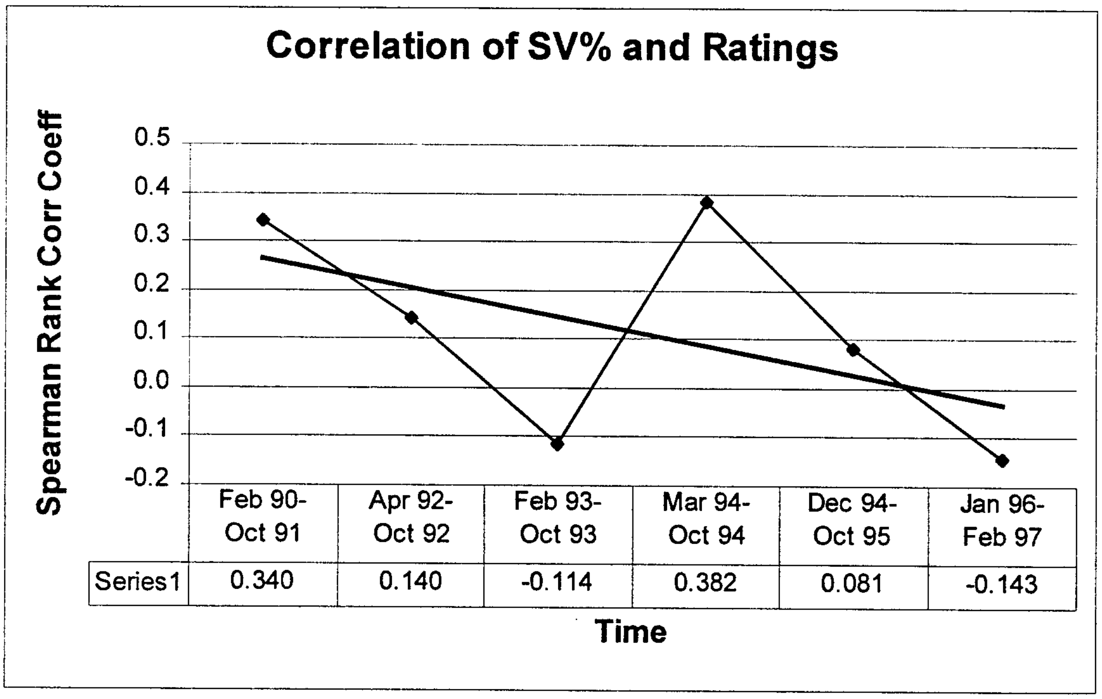

Figure 23. Cumulative SV% and Ratings Correlation Trend

Investigative Question #3

Do the past performance ratings correlate with measures of profitability? Are the perceived "best" contractors actually the most "profitable?" The relationship between profitability measures, CPARS color ratings, and objective contract performance measures will be examined through the remainder of the hypothesis tests.

Hypotheses 7-14

Hypotheses 7 and 8 explore the relationship between two measures of corporate profitability, Return on Equity percentage (ROE%) and Return on Investment percentage (ROI%), and CPARS cost color ratings. The next two hypotheses evaluate the profitability measures against period CV%. Again, to add to the robustness of this effort, the cumulative CV% was also used in the correlation tests. These hypotheses were named 11* and 12*. Because the CPARS schedule color ratings did not reject the null hypothesis and exhibit reliability, they will not be evaluated against profitability measures. Table 11 lists the Spearman and Pearson correlations and their respective p-values found for the Hypothesis tests.

Table 11. Correlations for Hypotheses 7, 8, 11, 12, 11* and 12*

	Variables	**Spearman's Correlation**	**p-value**	**Pearson's Correlation**	**p-value**
H7	Cost Color & ROE%	-0.045	0.367	-0.234	0.074
H8	Cost Color & ROI%	0.085	0.262	-0.172	0.192
H11	Period CV% & ROE%	0.066	0.388	0.012	0.957
H12	Period CV% & ROI%	0.028	0.452	0.023	0.922
H11*	Cum CV% & ROE%	0.068	0.330	-0.179	0.256
H12*	Cum CV% & ROI%	0.253	0.049	-0.081	0.612

From Table 11 it can be seen the only Spearman p-value below 0.10 is for cumulative CV% and ROI%. The analysis shows that there is no correlation between corporate level profitability measures and cost color ratings. Also, there is no correlation between profitability measures and period CV%. Only one of the six null hypotheses tested can be rejected. The analysis shows that there is strong support for a slight correlation between cumulative CV% and ROI%. The strong support, however, remains somewhat questionable because of the poor p-value associated with the Pearson's correlation. Potential reasons for this finding will be discussed further in Chapter V.

Hypothesis 15

The purpose of this hypothesis is to determine if the mean ROE% is different for the CPARS cost color ratings. Figure 24 shows that the ANOVA p-value strongly supports the rejection of the null hypothesis and several intervals do not contain zero. From the initial evaluation of the data, there is strong support that there is a difference between the ROE% means of at least two color ratings. However, the color that is different, Red, is based on three data points. This significantly weakens the result of this analysis. In fact, extreme observations, or outliers can be seen in Figures 25 and 26. Once these potential outliers were removed, the test provided different results.

```
Analysis of Variance for ROE%
Source      DF        SS        MS        F        P
Cost         3      1093       364     2.91    0.043
Error       55      6894       125
Total       58      7987
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level        N      Mean     StDev  --+---------+---------+---------+----
Blue        20     15.95      3.61   (---*---)
Green       19     15.04      3.64  (----*---)
Yellow      17     19.36     15.59      (---*----)
Red          3     34.54     35.49            (----------*----------)
                                    --+---------+---------+---------+----
Pooled StDev =     11.20               12        24        36        48

Intervals for (column level mean) - (row level mean)

              Blue        Green        Red
  Green      -8.60
             10.42

  Red       -36.97       -37.95
             -0.21        -1.06

  Yellow    -13.20       -14.23       -3.41
              6.39         5.59       33.78
```

Figure 24. Hypothesis 15 Tukey Intervals

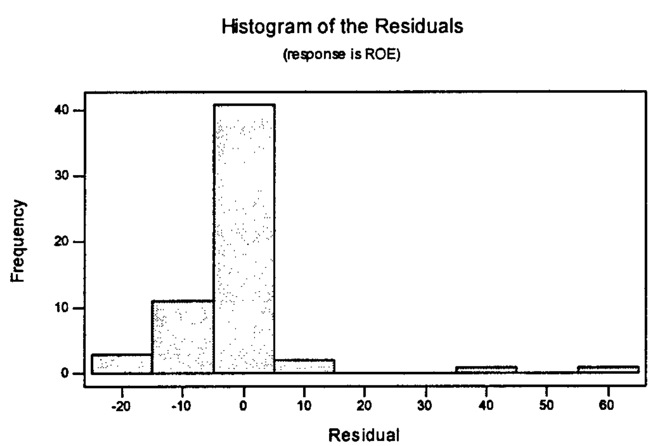

Figure 25. Hypothesis 15 Residual Histogram

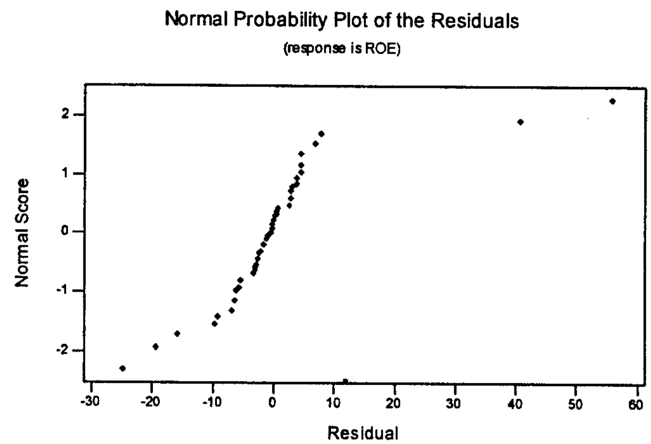

Figure 26. Hypothesis 15 Normal Probability Plot

After removing the outliers, the ANOVA p-value no longer supports the rejection of the null hypothesis. Figure 27 shows the Tukey test results for Hypothesis 15 after the removal of the outlier. Without the outliers, the Tukey multiple comparison procedure no longer shows a difference between the mean ROE% of the color ratings.

Thus, the result that the finding that there is a difference between the population means of cost color ratings when using ROE% is very questionable. The implications of this finding will be discussed further in the next chapter.

```
Analysis of Variance for ROE%
Source      DF        SS         MS         F        P
Cost         3      13.4        4.5      0.21    0.886
Error       53    1103.8       20.8
Total       56    1117.1
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level        N     Mean     StDev  --------+---------+---------+--------
Blue        20   15.951     3.609                    (-----*----)
Green       19   15.038     3.641                 (-----*-----)
Yellow      16   15.869     6.198                   (-----*------)
Red          2   14.220     6.435  (------------------*-----------------)
                                   --------+---------+---------+--------
Pooled StDev =    4.564                  10.5      14.0      17.5

Intervals for (column level mean) - (row level mean)

              Blue        Green        Red
Green       -2.964
             4.790

Red         -7.243       -8.177
            10.706        9.814

Yellow      -3.977       -4.937     -10.725
             4.141        3.275       7.426
```

Figure 27. Hypothesis 15 Tukey Intervals (Outliers Removed)

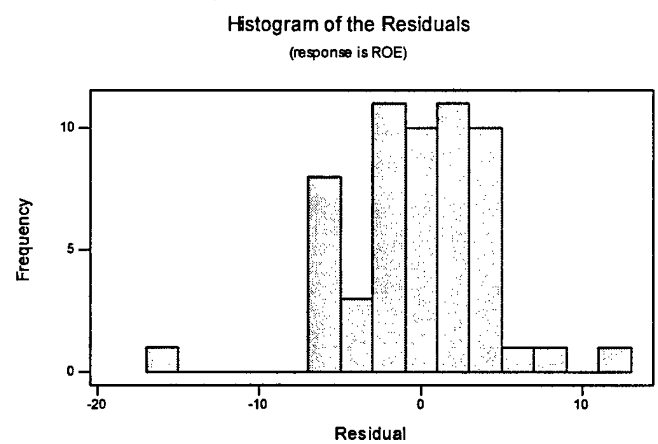

Figure 28. Hypothesis 15 Residual Histogram (Outliers Removed)

Hypothesis 16

The purpose of this hypothesis is to determine if the mean ROI% is different for the CPARS cost color ratings. Figure 29 shows that the ANOVA p-value moderately supports the rejection of the null hypothesis yet all intervals still contain zero. From the initial evaluation of the data, there is moderate support that there is a difference between the ROI% means of at least two color ratings. However, the color that is different, Red, is again based on only three data points. This significantly weakens the result of this analysis. In fact, extreme observations, or outliers can be seen in Figures 30 and 31. Just as the test with ROE%, once these potential outliers were removed, the test provided different results.

```
Analysis of Variance for ROI%
Source     DF        SS          MS         F        P
Cost        3     195.3        65.1      2.20    0.098
Error      55    1626.6        29.6
Total      58    1821.9
                                   Individual 95% CIs For Mean
                                   Based on Pooled StDev
Level       N      Mean     StDev  -----+---------+---------+---------+-
Blue       20     6.091     2.511   (----*----)
Green      19     5.315     2.071  (----*----)
Yellow     17     6.952     7.264     (----*----)
Red         3    13.843    17.107          (------------*-----------)
                                   -----+---------+---------+---------+-
Pooled StDev =    5.438                5.0      10.0      15.0      20.0

Intervals for (column level mean) - (row level mean)

              Blue        Green        Red
 Green      -3.843
             5.396

 Red       -16.681      -17.487
             1.176        0.430

 Yellow     -5.618       -6.451     -2.139
             3.896        3.177     15.922
```

Figure 29. Hypothesis 16 Tukey Intervals

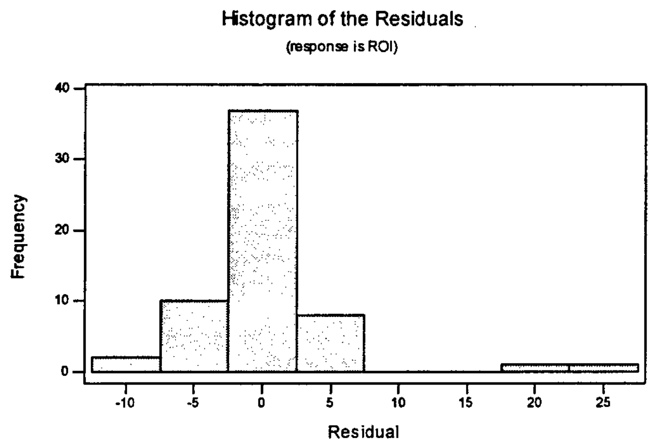

Figure 30. Hypothesis 16 Residual Histogram

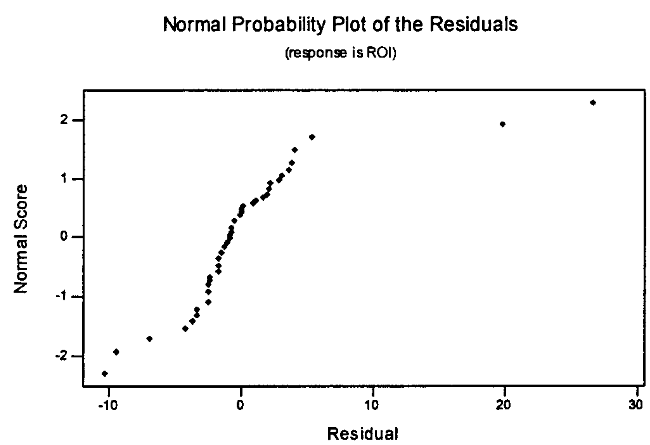

Figure 31. Hypothesis 16 Normal Probability Plot

After removing the outliers, the Tukey p-value no longer supports the rejection of the null hypothesis. Figure 32 shows the Tukey test results for Hypothesis 16 after the removal of the outlier. Without the outliers, the Tukey multiple comparison procedure no longer shows a difference between the mean ROI% of the color ratings.

Thus, the result that the finding that there is a difference between the population means of cost color ratings when using ROI% is very questionable. The implications of this finding will be discussed further in the next chapter.

```
Analysis of Variance for ROI%
Source      DF        SS         MS         F        P
Cost         3     12.99       4.33      0.80    0.501
Error       53    287.75       5.43
Total       56    300.74
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level        N      Mean     StDev  -------+---------+---------+---------
Blue        20     6.091     2.511                        (----*-----)
Green       19     5.315     2.071                    (-----*----)
Yellow      16     5.287     2.455                    (----*-----)
Red          2     3.970     0.608  (---------------*---------------)
                                    -------+---------+---------+---------
Pooled StDev =     2.330                  2.0       4.0       6.0

Intervals for (column level mean) - (row level mean)

             Blue         Green        Red
 Green      -1.203
             2.756

 Red        -2.461       -3.248
             6.703        5.938

 Yellow     -1.268       -2.069       -5.951
             2.876        2.124        3.317
```

Figure 32. Hypothesis 16 Tukey Intervals (Outliers Removed)

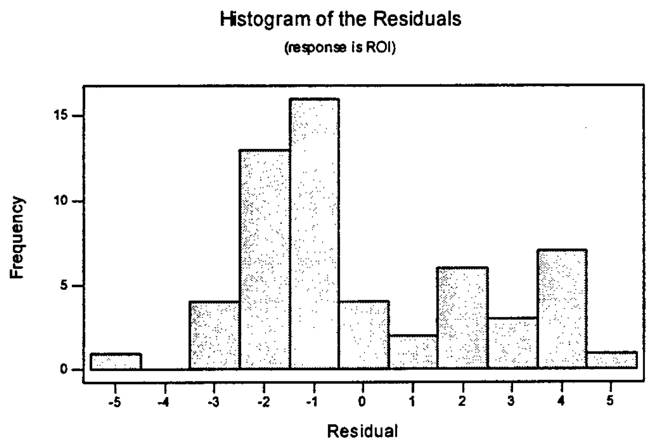

Figure 33. Hypothesis 16 Residual Histogram (Outliers Removed)

Summary

The results of the 26 hypotheses are summarized in Table 12. The Spearman rank correlation coefficient was the primary evaluation technique and was supplemented by the Pearson's product moment correlation for ten of the hypotheses. Tukey's multiple comparison technique was implemented for six of the tests. Regression was performed on Spearman correlations over time for four of the hypothesis tests. The other six hypotheses were not tested because the correlation of schedule color ratings and objective measures of performance was deemed inconsequential.

The hypothesis tests of the first two Investigative Questions provide several noteworthy results. First, there is only moderate support for a slight correlation between the period CV% and cost color ratings. Also, the period SV% and schedule color rating has improved over time. Cumulative CV%, however, appears to be the primary determinant of cost color ratings and its relationship to the ratings has diminished over time.

Table 12. Summary of Hypothesis Tests

		Result	Comment
IQ #1	H1	Reject	Moderate support
	H2	Fail to reject	Negative correlation
	H3	Fail to reject	Tukey p-value of 0.170
	H4	Reject	Moderate support
IQ #1*	H1*	Reject	Strong support
	H2*	Fail to reject	Spearman p-value of 0.106
	H3*	Reject	Strong support
	H4*	Fail to reject	Tukey p-value of 0.893
IQ #2	H5	Fail to reject	Period CV% has not changed
	H6	Reject	Strong support
IQ #2*	H5*	Reject	Strong support
	H6*	Fail to reject	Regression p-value of 0.303
IQ #3	H7	Fail to reject	Spearman p-value of 0.367
	H8	Fail to reject	Spearman p-value of 0.262
	H9	N/A	Failed to reject H2 & H2* H_0
	H10	N/A	Failed to reject H2 & H2* H_0
	H11	Fail to reject	Spearman p-value of 0.388
	H12	Fail to reject	Spearman p-value of 0.452
	H13	N/A	Failed to reject H2 & H2* H_0
	H14	N/A	Failed to reject H2 & H2* H_0
	H15	Reject	Strong support – questionable
	H16	Reject	Moderate support – questionable
	H17	N/A	Failed to reject H2 & H2* H_0
	H18	N/A	Failed to reject H2 & H2* H_0
IQ #3*	H11*	Fail to reject	Spearman p-value of 0.330
	H12*	Reject	Strong support with Spearman

The third Investigative Question did not provide as astounding results. Cumulative CV% did show strong support for a slight correlation with ROI% using the Spearman correlation. However, the Pearson correlation was found to be highly insignificant between cumulative CV% and ROI%. Another minor finding was that the mean period ROE% was shown to be different for the Red color rating. However, this finding remains questionable since it was based on a sample size of only three Red ratings.

Table 13 lists the strongly supported findings from this effort. The first strongly supported finding is that cumulative CV% is moderately correlated with cost color ratings. Second, the mean cumulative CV% is different for at least one color rating. Furthermore, only the adjacent color ratings were overlapping. Third, the reliability of period SV% has significantly improved over time. Finally, the relationship between cumulative CV% and cost color ratings has diminished over time. Potential reasons for and the ramifications of these findings will be discussed more thoroughly in the final chapter.

Table 13. Summary of Strongly Supported Findings

IQ #	Parameters	Comment
1*	Cumulative CV% and rating	Strong support for correlation
1*	Cumulative CV% and rating	Average ratings are different
2	Period SV% and rating	Relationship is improving
2*	Cumulative CV% and rating	Relationship is weakening

V. Conclusions and Recommendations

Overview

This chapter provides the conclusions from the analysis presented in Chapter IV and discusses the findings of this research effort. All conclusions discussed in this section refer to the tests evaluated without outliers, except where indicated. Additionally, recommendations are made to improve the reliability of the CPARS process and also for future research. Table 14 summarizes the primary conclusions that will be discussed further throughout this chapter.

Table 14. Primary Conclusions

1	*Cumulative* cost performance measures are a primary determinant of *period* cost color ratings and do discriminate between contractor performances
2	The reliability of cumulative cost performance measures and cost color ratings has significantly weakened over time
3	Period schedule performance measures are not yet a significant aggregate determinant of schedule color ratings, but its reliability has improved significantly over time

CPARS Ratings and Reliability

Cost Performance Ratings

As discussed earlier, CPARS policy requires that period color ratings be based on objective measures such as Cost Variance (CV) or Cost Variance Percentage (CV%).

Also, the rating must be based on performance *during* that period. Thus, logic dictates that the period CV% would be the primary determinant of the cost color ratings.

As shown in Chapter IV, there is moderate support that there is only a slight correlation between period cost measures and the cost color ratings. This result is surprising due to the fact that the aforementioned CPARS policy explicitly states that the report should contain period performance evaluations only. A simple calculation of CPR or C/SSR data by the program offices would yield a period variance and period variance percentage. The evaluator could objectively evaluate the contractor's cost (and schedule) performance during the period. The objective rating could then be anchored by the period cost measures and adjusted for any other objective or subjective information known at the time of rating.

In accordance with the above finding, the data analysis suggests that the mean period cost measures are not different for cost color ratings. If the period cost measures are not the primary determinant of the cost color ratings, then any differentiation found in the color ratings when using period cost measures is simply coincidental.

The implication of these results is that there may be confusion with the policy spelled out in AFMCI 64-107. It appears that evaluators may be relying on cumulative cost measures instead of period objective measures when assigning ratings.

Cumulative cost measures, on the other hand, are a significant indicator of what color rating a contractor will receive for their performance during the given period. There is strong support that there is a moderate correlation between cumulative cost measures and cost color ratings. In fact, these correlations ($\rho = 0.524$ for Spearman and $\rho = 0.447$

for Pearson) were the highest aggregate correlations encountered during this study. The policy of having the cumulative CV% data reported verses the period CV% data may actually reinforce this phenomenon.

Also, the cost color ratings do delineate between contractors' performance using cumulative cost measures. The analysis strongly supports that the mean cumulative cost measures are different for at least two color ratings. Also, adjacent color ratings did overlap, but non-adjacent color ratings did not overlap. A Blue rating, for example, does provide distinction between the cumulative performance of Yellow and Red ratings. However, due to overlap, a Green rating does not necessarily discriminate between Blue or Yellow ratings. In other words, a contractor's cumulative objective measurement of cost performance, namely CV%, provides the basis of the color rating, which will be Green for this example. Other objective or subjective information determines whether the color rating remains the same or is changed to a Blue or Yellow. This distinction should provide value during source selection evaluations because the ratings do discriminate between non-adjacent cost color ratings. This result, that cumulative cost performance measures are a primary determinant of cost color ratings and do discriminate between performances, is the first primary conclusion of this thesis.

A recommendation, then, to improve CPARS involves the policy concerning the cost ratings. AFMC should either request a cumulative cost color rating or request *period* cost and schedule variance percentages instead of cumulative cost and schedule variance percentages. The solution to this question can only be found by answering underlying questions. An example of these questions includes, "Do we want to select contractors that performed well during a given percentage of arbitrary periods, or over the entire effort?" The AF must first determine which information would be more beneficial during a source selection before choosing an alternative. Nevertheless, the bottom line is that AFMC must either change policy or alter training to ensure raters understand what is being evaluated.

Schedule Performance Ratings

As shown in Chapter IV, there is no support for any correlation between period schedule measures and the schedule color ratings at an aggregate level. As with the period cost measures and cost color ratings, this result is surprisingly in contrast with CPARS policy. Again, a simple calculation of CPR or C/SSR data by the program offices would yield a period variance and period variance percentage. The evaluator could objectively evaluate the contractor's schedule (and cost) performance during the period. The objective rating could then be anchored by the period schedule measures and adjusted for any other objective or subjective information known at the time of rating.

In accordance with the finding that period schedule measures are not a primary determinant of schedule color ratings, the data analysis suggests that the mean period schedule measures are not different for schedule color ratings. As with period cost measures, if the period schedule measures are not the primary determinant of the schedule color ratings, then any differentiation found in the color ratings when using period schedule measures is coincidental.

As with period cost measures, the implication of these results is that evaluators are not following the policy spelled out in AFMCI 64-107. Precisely, the CPARS schedule color ratings are not yet based on objective facts for the period evaluated.

Unlike cumulative cost measures, Chapter IV shows that there is no support for any correlation between cumulative schedule measures and the schedule color ratings at an aggregate level. Reporting cumulative SV% with the color ratings does not seem to reinforce using the cumulative SV% as a basis for the color ratings as it does with cumulative CV%. Other objective or subjective information must be responsible for determining the schedule color ratings.

Not surprisingly, cumulative schedule measures do not discriminate between different color ratings. As with period schedule measures, the data analysis suggests that the mean cumulative schedule measures are not different for schedule color ratings. Again, if the cumulative schedule measures are not a primary determinant of the schedule color ratings, then any differentiation found in the color ratings when using cumulative schedule measures is coincidental.

Thus, CPARS schedule color ratings do not yet correlate with period or cumulative objective measures. Other objective or subjective factors not included in this study provide the basis for the ratings. Because neither period schedule measures nor cumulative schedule measures were proven to be reliable in determining the period color rating, they were not evaluated against profitability measures.

A second suggestion to improve the reliability of CPARS then relates to the schedule color ratings. Currently, the objective measure, SV%, is not being used as a primary determinant of schedule color ratings. Either the use of period SV% in determining schedule color ratings needs to be reemphasized, or different objective measures for assessing schedule performance need to be identified and presented to raters as options to use as a basis for developing ratings. Because the color ratings must be based on objective measures, any new measures need to be identified and made available to raters. It is also recommended that source selection evaluators use another discriminating factor until the reliability of the schedule color rating and objective period schedule measures improve.

CPARS Reliability vs. Time

There is no support that the reliability of the CPARS color ratings with respect to period cost measures has changed over time. As discussed earlier, there is only a slight correlation between period cost measures and cost color ratings. If the period cost measures are not the primary determinant of the cost color rating, then any change over time must be purely coincidental.

The relationship between cumulative cost measures and cost color ratings, on the other hand, has changed significantly over time. In fact, the analysis in Chapter IV shows that there is strong support that the relationship has weakened over time. In short, the reliability of the CPARS using cumulative cost measures is weakening over time. This is the second primary conclusion of this effort. If the USAF truly wants past performance ratings to be based on period measures and if the reliability of past performance ratings with those period measures were improving, then this occurrence would be desirable.

The analysis in Chapter IV shows that cumulative cost measures were once a strong discriminator of a contractor's performance. The relationship now between the objective cumulative cost measures and cost color ratings is significantly weakening. One possible explanation for this decline in reliability can be tied to a change in acquisition policy. The decline began roughly during CY1994, immediately prior to DoD issuing guidance implementing Integrated Product Teams (IPTs) through the entire acquisition process. Now, with the "Team" viewpoint, a poor contractor grade implies a poor performance by the evaluator as well. Evaluators are now in the position of rating themselves, not just the contractor. Because people are often hesitant to report their own performance as being poor, the reliability of the CPARS ratings appears to be diminishing.

A third suggestion to improve CPARS reliability would be to evaluate the consequences of policy on CPARS ratings. This evaluation would preferably take place before implementation of any PPI policy. Negative impacts must be explored and minimized. This topic is also discussed in the Recommendations for Further Research section.

The third primary conclusion of this research is that, even though the period schedule measures are not a determinant of the schedule color rating, the correlation between the two has changed over time. The data strongly supports that the correlation has not only changed, but it has improved over time. Therefore, despite the weak relationship of schedule color ratings and period schedule measures in an aggregate sense, their reliability has significantly improved over time.

This result provides a different picture from the previous finding that both period and cumulative schedule measures were not primary determinants of schedule color ratings. This finding does indicate that evaluators are beginning to use objective measures, such as period SV%, to begin the color rating determination.

Conversely, there is no support that the reliability of the CPARS schedule color ratings with respect to cumulative schedule measures has changed over time. As discussed earlier, there is no correlation between cumulative schedule measures and schedule color ratings. Further, if the cumulative schedule measures are not the primary determinant of the cost color rating, then any change over time would be purely incidental.

Cost Measures and Profitability Measures

There is no support that the cost color rating is correlated with profitability measures such as ROE% and ROI%. A possible rationale for this could be that Industry has shielded itself to the impacts of a reduced DoD budget. By restructuring, Industry has minimized the effects of the budget reductions, and therefore, the use of PPI. An alternate rationale could be that CPARS is still lacking as a performance discriminating mechanism. Recall the Limitations provided in Chapter III gives another reason for a lack of correlation. This reason is that the profitability measures are corporate measures and cost measures are for single contracts.

As with the cost color ratings, there is no correlation between period cost performance measures and corporate profitability measures. The cumulative cost performance measures, however, displayed some correlation with respect to profitability measures. Although cumulative cost performance measures and one of the profitability measures showed strong support for a slight positive correlation, no grand inferences can be made to this result. The crosscheck using the Pearson's correlation did not show any significant correlation. Even so, this finding of slight correlation does provide insight for future research of this area.

Recommendations for Further Research

The focus of any follow-on research needs to explore the relationship of IPT implementation and the weakening of the cost color rating reliability. Are raters actually being put in the position of rating themselves? If so, can the raters sacrifice personal biases and egos to provide a truly effective evaluation that can discriminate between Marginal, Satisfactory, and Very Good performances?

Since objective measures, such as SV%, are not primary determinants for developing schedule color ratings, new metrics must be developed. These metrics must be actual discriminators of past performance. Examples of this research would be to develop indicators of proactive or reactive management with respect to "unknown-unknowns" and also how can DoD objectively measure these indicators.

A related topic for future research would concern technical performance. What objective measures exist and have the best correlation with technical performance parameters? In other words, what quantifiable measures can be discriminators of actual technical performance?

Another recommendation is to categorize ratings by CAGE Code to identify if CPARS should be even more specific. "More specific" in this sense means not just evaluate the contractor on similar efforts, but evaluate the CAGE Code that will be performing the bulk of the work and their similar efforts.

A fifth recommendation is to evaluate the investigative questions of this effort using the entire AFMC CPARS database. This will either strengthen or refute the results of this effort. It could also provide insights to any policy or process differences between the acquisition centers where the CPARS are stored.

A final recommendation concerns the impact of recent corporate acquisitions and mergers on past performance history. What impact will there be on the AF's ability to use this process in the near horizon until more PPI data on the restructured corporations can be obtained? Any research conducted in this area can further examination the slight correlation between cumulative CV% and corporate profit measures.

Summary

DoD is attempting to capitalize on the Industry trend of establishing long-term relationships with reliable suppliers. One of the criteria Industry uses to pick these "reliable suppliers" is past performance. The Department of Defense is also using past performance as an evaluation factor in source selections. Air Force Material Command (AFMC) employs the Contractor Performance Assessment Reporting System (CPARS).

Cumulative cost performance measures were once a strong discriminator of a contractor's cost rating. Yet, the relationship between the objective cumulative cost performance measures and cost color ratings has begun to weaken. One possible explanation for this decline in reliability can be traced to changes in acquisition policy. With the implementation of IPTs, evaluators are now implicitly appraising themselves as well as the contractor.

Also, the period schedule performance measures are not a significant factor in determining the contractor's rating in aggregate. Nonetheless, the correlation of period schedule performance measures and the schedule color rating has improved over time. Perhaps with additional training, the period schedule performance measures may become a strong determinant of the schedule color rating.

The last question to be answered is, “if period cost and schedule measures are not determinants of color ratings, then why does AFMC policy order that they are used?” The answer lies with the fact that these objective measures are based on “planned” work. The period objective measures then are used to evaluate the contractors’ performance based on that plan. Thus, the CPARS color ratings provide source selection officials information about the contractor's performance to their plans on previous efforts. Finally, if decisions in source selections require PPI discerning the contractor’s performance to their plans, then period objective cost and schedule performances are the best measures to base CPAR ratings.

Next, there is no relationship between cost color ratings and measures of profitability. Industry firms seem to have insulated themselves to the instability inherent in DoD acquisition.

Although CPARS policy mandates evaluations based on using period performance, the cost color ratings are more related to cumulative performance. Thus, the author recommends that AFMC either change CPARS cost rating policy to reflect the use of cumulative objective measures or provide additional training so evaluators better understand what is assessed during a CPARS rating period.

Appendix: Data Tables.

Cage Code	Contract Number	Begin Period	End Period	Percent Complete	Cost Rating	Sched Rating	CPARS CV %	CPARS SV %	Period CV %	Period SV %
1	A	14-Jun-96	15-Feb-98	55.00	Green	Green	6.10	71.13		
1	B	14-Feb-96	14-Feb-97	65.00	Yellow	Yellow	0.31	1.39		
1	B	15-Feb-97	14-Feb-98	81.00	Red	Red	-149.50	1.39		
2	C	16-Feb-96	16-Feb-97	50.00	Yellow	Blue	-27.00	5.10		
3	D	1-Jan-96	31-May-97	75.00	Yellow	Yellow	-11.00	-14.00		
4	E	1-Sep-93	31-Aug-94	75.00	Green	Green				
4	E	1-Sep-94	31-Jan-96	100.00	N/A	Green				
5	F	20-Mar-96	31-Jan-97	23.00	Green	Yellow	N/A	N/A		
5	F	1-Feb-97	31-Jan-98	39.00	Yellow	Green	-39.00	-5.00		
5	G	6-Aug-91	20-Jun-92	20.00	Green	Green	31.10	0.00		
5	G	20-Jun-92	20-Jun-93	29.00	Green	Green		0.00		
5	G	1-Jun-93	31-May-94	50.00	Blue	Green	7.40	-4.10		
5	G	1-Jun-94	31-May-95	64.00	Blue	Green	3.90	-1.20		
5	G	1-Jun-95	31-May-96	86.00	Blue	Blue	9.00	0.00		
5	G	1-Jun-96	31-May-97	100.00	Blue	Blue	6.00	0.00		
6	H	2-Apr-92	31-Jan-93	51.00	Blue	Blue	8.30	-3.70	14.22	-5.08
6	H	1-Feb-93	31-Jan-94	99.00	Blue	Blue	5.30	-0.80	1.60	3.48
6	H	1-Feb-94	28-Feb-95	100.00	Blue	Blue	-5.01	0.00		
7	I	25-Aug-93	31-May-94	20.00	Yellow	Green	21.60	0.00		
7	I	1-Jun-94	31-May-95	50.00	Yellow	Green	0.00	3.00		
7	I	1-Jun-95	31-May-96	70.00	Yellow	Yellow	0.10	-0.80		
7	I	1-Jun-96	31-May-97	85.00	Green	Green	6.00	0.20		
8	J	1-Jan-94	30-Dec-94	75.00	Yellow	Green	10.00			
8	J	1-Jan-95	31-Dec-95	100.00	Green	Green				
9	K	1-Oct-90	30-Sep-91	32.00	Red	Green	-29.00	-11.00		
9	K	1-Oct-91	30-Sep-92	64.00	Red	Green	-31.00	-15.00		
9	K	1-Oct-92	30-Sep-93	63.00	Yellow	Yellow	1.16	-1.47		
9	K	1-Oct-93	30-Sep-94	68.00	Yellow	Yellow	0.20	-0.20		
9	K	1-Oct-94	03-Sep-95	87.00	Red	Red	-2.30	-2.20		
9	K	4-Sep-95	30-Sep-96	92.00	Red	Yellow	-4.69	-1.69		
9	K	1-Oct-96	30-Sep-97	99.90	Red	Green	N/A	N/A		
10	L	18-Mar-91	30-Jun-92	26.00	Yellow	Yellow	-11.70	-4.60	13.17	1.24
10	L	1-Jul-92	30-Apr-93	48.00	Red	Red	-25.13	-5.10	11.91	0.00
10	L	1-May-93	28-Feb-94	53.00	Yellow	Yellow	-35.00	-6.00	3023.32	-68.08
10	L	1-Mar-94	28-Feb-95	62.00	Green	Green	-0.60	-0.90		
10	L	1-Mar-95	08-Jul-95	66.00	Green	Green	-1.20	-0.80		
11	M	1-Jun-96	31-May-97	94.00	Green	Green	4.00	1.00		
12	N	31-May-92	31-May-93	42.00	Yellow	Yellow	-14.80	-8.80	-12.14	6.18
12	N	31-May-93	31-May-94	73.00	Yellow	Green	-27.90	-9.10	-21.92	1.22
12	N	1-Jun-94	31-May-95	99.00	Red	Green	15.68	N/A		
12	N	1-Jun-95	31-May-96	99.00	N/A	Red	N/A	N/A		
13	O	1-Jun-96	31-May-97	91.00	Blue	Blue	-0.05	-0.05		
14	P	1-Jan-94	30-Dec-94	95.00	Blue	Green				
15	Q	1-Jul-94	30-Jun-95	88.00	Blue	Green	N/A	-1.00	2.70	-2.66
15	Q	1-Jul-95	30-Jun-96	81.40	Green	Green	N/A	-1.00	-4.63	5.42
15	Q	1-Jul-96	17-Jun-97	99.00	Green	Green	N/A	-1.00	10.16	2.22
15	R	1-Feb-95	31-Jan-96	28.00	Green	Green	0.70	-0.40	0.06	-0.18
15	R	1-Feb-96	14-Feb-97	60.20	Blue	Green	1.34	-1.81	-1.56	-1.97
15	S	21-Apr-95	20-Apr-96	45.00	Blue	Blue	3.60	-3.40		
15	S	21-Apr-96	20-Apr-97	87.20	Blue	Blue	-4.20	-0.60		

15	S	21-Apr-97	20-Apr-98	88.00	Blue	Blue	-3.50	-0.02		
15	T	1-Feb-97	31-Jan-98	26.52	Blue	Blue	N/A	N/A		
16	U	1-Oct-92	30-Sep-93	37.00	Green	Yellow	-3.00	-2.96	-3.49	6.64
16	U	1-Oct-93	30-Sep-94	45.00	Green	Green	-3.37	-1.08	-4.35	5.01
16	U	1-Oct-94	30-Sep-95	54.00	Yellow	Green	-3.80	-2.30	8.65	0.39
16	U	1-Oct-95	30-Sep-96	70.10	Yellow	Green	-2.60	-1.10	-15.88	-3.69
16	U	1-Oct-96	30-Sep-97	73.40	Yellow	Green	0.10	-0.50	14.28	3.12
16	V	1-Jan-95	31-Dec-95	100.00	Blue	Green	N/A	N/A		
16	W	1-Jan-96	30-Jun-97	75.00	Blue	Green	-0.20	-6.40		
17	X	8-Jun-94	30-Sep-95	45.00	Green	Green	-10.00	-35.00		
18	Y	21-Jun-94	20-Jun-95	41.00	Yellow	Yellow	-12.70	-31.90		
18	Y	21-Jun-95	30-Sep-96	78.30	Blue	Blue	5.60	-3.60		
18	Y	1-Oct-96	30-Sep-97	82.40	Green	Green	N/A	N/A		
19	Z	1-Jun-94	31-May-95	50.00	Green	Green	6.58	-4.35		
19	Z	1-Jun-95	31-May-96	95.00	Blue	Green	6.58	-4.35		
19	AA	1-Jun-96	31-May-97	67.00	Green	Blue	5.53	-5.11		
19	AB	27-Sep-91	31-May-92	17.00	Blue	Green	30.34	1.35		
19	AB	1-Jun-92	31-May-93	38.00	Blue	Green	25.01	15.52		
19	AB	1-Jun-93	31-May-94	75.00	Green	Green	22.00	-17.00	26.74	-26.06
19	AB	1-Jun-94	31-May-95	90.00	Green	Green	-11.00	-21.00	-32.69	-314.52
19	AB	1-Jun-95	31-May-96	87.00	Blue	Green	14.60	-4.20	-27.30	-30.16
19	AB	1-Jun-96	31-May-97	92.00	Blue	Green	9.90	-9.30	-4.09	60.96
19	AB	1-Jun-97	31-Dec-97	100.00	Blue	Green	6.30	0.00		
19	AC	23-Sep-92	31-May-93	30.00	Green	Green	-9.88	-33.30		
19	AC	1-Jun-93	31-May-94	□	Green	Green	-39.00	-24.00		
19	AC	1-Jun-94	31-May-95	82.00	Green	Green	-39.82	-12.67	-50.96	40.29
19	AC	1-Jun-95	31-May-96	100.00	Green	Green	N/A	N/A		
19	AD	23-Mar-93	30-May-94	22.00	Blue	Green	26.90	16.20	26.94	-16.17
19	AD	1-Jun-94	31-May-95	34.00	Blue	Yellow	16.78	-12.73	-0.28	-4.21
19	AD	1-Jun-95	31-May-96	53.00	Green	Yellow	5.70	-6.60	-34.50	22.41
19	AD	1-Jun-96	31-May-97	69.80	Green	Green	0.80	-18.90	-49.80	-60.29
19	AD	1-Jun-97	31-Dec-97	100.00	Green	Green	0.00	0.00	1.32	-63.38
19	AE	30-Sep-92	31-May-93	20.00	Green	Green	19.72	13.09	6.32	-9.29
19	AE	1-Jun-93	31-May-94	63.00	Green	Green	0.63	-9.80	9.58	-11.63
19	AE	1-Jun-94	31-May-95	95.00	Green	Green	8.79	-1.06	11.87	77.24
19	AF	1-Jun-96	31-May-97	93.00	Blue	Yellow	N/A	N/A	3.05	-35.52
19	AG	16-Jul-96	31-May-97	88.00	Green	Green	0.20	-1.00		
19	AH	1-Jun-94	31-May-95	16.00	Blue	Blue	-0.74	-2.39	7.72	7.38
19	AH	1-Jun-95	31-May-96	50.00	Green	Blue	-5.30	-3.10	-8.15	-3.04
19	AH	1-Jun-96	31-May-97	80.00	Green	Blue	-5.64	-0.38	-3.29	3.38
19	AI	31-Jul-96	30-Sep-97	25.70	Blue	Green	10.40	-5.70		
19	AJ	2-Sep-96	31-May-97	20.00	Blue	Blue	18.00	-38.00		
19	AK	16-Oct-96	27-Jun-97	41.00	Blue	Green	0.11	-3.76		
19	AL	31-Jan-97	30-Jan-98	19.00	Blue	Green	N/A	N/A	7.31	-8.27
20	AM	1-Jun-93	31-May-94	77.60	Green	Green	1.00	-3.50		
20	AM	1-Jun-94	31-May-95	96.00	Green	Green	2.76	-5.30		
20	AM	1-Sep-94	31-May-95	26.00	Green	Green	-9.90	-5.30	-5.97	-4.69
20	AM	1-Jun-95	31-May-96	85.00	Green	Green	-3.30	-1.00	-1.46	2.50
20	AM	1-Jun-96	31-May-97	97.00	Blue	Green	1.30	0.00	23.20	1.94
21	AN	2-Feb-90	01-Feb-91	34.00	Yellow	Yellow	-11.09	-28.47		

21	AN	2-Feb-91	01-Feb-92	48.00	Green	Green	-16.89	-45.96		
21	AN	1-Feb-93	01-Feb-94	79.00	Yellow	Green	-12.20	-1.10		
21	AN	1-Feb-94	01-Feb-95	88.00	Yellow	Green	-11.00	-9.00		
21	AN	1-Feb-95	01-Feb-96	98.00	Yellow	Yellow	-11.00	-0.70		
21	AN	2-Feb-97	01-Feb-98		Green	Yellow				
21	AO	1-Jun-92	31-May-93	20.00	Yellow	Yellow	N/A	N/A	9.38	-3.68
21	AO	1-Jun-93	31-May-94	42.00	Red	Green	N/A	N/A	1.07	-3.42
21	AO	1-Jun-94	31-May-95	64.00	Green	Green	7.70	7.50	-9.09	-19.86
21	AO	1-Jun-95	31-May-96	83.00	Blue	Green	0.00	-1.70		
21	AO	1-Jun-96	31-May-97	92.80	Blue	Blue	-6.10	-3.00		
21	AP	1-Jun-95	31-May-96	29.40	Blue	Green	15.30	3.00	14.78	-2.17
21	AP	1-Jun-96	31-May-97	100.00	Blue	Green	18.58	0.00	26.39	-1.14
21	AQ	12-Dec-94	31-May-95	18.00	Blue	Yellow	5.70	-8.80	6.45	-6.95
21	AQ	1-Jun-95	31-May-96	59.40	Blue	Green	13.50	-3.00	13.57	-1.64
21	AQ	1-Jun-96	31-May-97	100.00	Blue	Blue	8.50	N/A	-0.29	3.50
21	AR	1-Jun-96	31-May-97	60.00	Blue	Green	N/A	N/A		
22	AS	1-Jan-95	30-Jan-96	100.00	Blue	Green				
22	AT	3-Jun-96	31-Jan-97	15.00	Yellow	Yellow	N/A	N/A		
23	AU	1-Jul-95	30-Sep-96	12.00	Green	Green	5.80	13.90		
23	AU	1-Oct-96	30-Sep-97	68.20	Green	Blue	-0.95	-4.10		
24	AV	8-Jun-95	30-Sep-96	88.60	Yellow	Yellow	-3.30	N/A		
25	AW	1-Sep-88	01-Feb-90	90.00	Green	Green	1.00	0.00	1.18	-13.11
25	AW	2-Feb-90	31-Jan-91	34.00	Yellow	Yellow	-11.50	-11.80	-27.28	-11.09
25	AW	1-Feb-91	01-Feb-94	82.00	Yellow	Yellow	-11.20	-2.90		
25	AW	1-Feb-94	01-Feb-95	87.00	Yellow	Yellow	-12.00	-2.40		
25	AW	2-Feb-95	01-Feb-96	98.00	Yellow	Yellow	-12.00	-2.40		
26	AX	1-Feb-95	31-Jan-96	0.00	Green	Blue	<3	1.00		
26	AX	1-Feb-96	31-Jan-97	90.00	Green	Yellow	<5	<5		
27	AY	12-Jan-97	31-Jan-98	25.57	Red	Yellow	39.09	39.48		
28	AZ	29-Apr-92	28-Apr-93	31.00	Yellow	Green	-20.00	-15.70	-24.01	-27.58
28	AZ	29-Apr-93	28-Apr-94	68.00	Red	Green	-42.00	-19.00	-55.09	-14.06
28	AZ	29-Apr-94	28-Apr-95	80.00	Red	Yellow	-90.00	-15.00	33.04	19.32
29	AAA	1-Jun-94	31-May-95	90.00	Blue	Green	7.50	-1.10		
29	AAA	1-Jun-95	31-May-96	98.00	Blue	Green	0.10	0.90		
29	AAA	1-Jun-96	31-May-97	100.00	Blue	Green	0.10	0.90		
30	AAB	1-Aug-91	30-Sep-92	10.00	Yellow	Green	-2.26	-6.86	-2.22	-6.87
30	AAB	1-Oct-92	30-Sep-93	18.00	Green	Green	-0.50	-2.00	0.63	1.00
30	AAB	1-Oct-93	30-Sep-94	33.00	Green	Green	-2.70	-1.50	-5.96	-0.35
30	AAB	1-Oct-94	30-Sep-95	43.60	Green	Green	-0.10	-0.70	5.06	0.44
30	AAB	1-Oct-95	30-Sep-96	57.90	Yellow	Green	-3.10	-1.50	-14.38	-2.99
30	AAB	1-Oct-96	30-Sep-97	68.70	Yellow	Green	0.20	-0.20	14.27	3.44
31	AAC	1-Oct-96	30-Sep-97	37.00	Green	Green	3.45	3.95		
32	AAD	16-Jun-91	15-Jun-92	77.00	Yellow	Blue	-8.69	-4.19		
33	AAE	26-Mar-96	31-Dec-97	52.00	Red	Red	N/A	N/A		
34	AAF	1-Apr-95	31-Mar-96	20.00	Yellow	Green	N/A	N/A		
34	AAF	1-Apr-96	31-Oct-96	35.00	Yellow	Green	N/A	N/A		
34	AAF	1-Nov-96	31-Oct-97	50.00	Green	Green	N/A	N/A		
35	AAG	31-May-96	31-May-97	74.00	Yellow	Yellow	12.10	N/A		
36	AAH	1-Oct-96	30-Sep-97	66.00	Green	Blue	9.60	8.00		

Bibliography

Aeronautical Systems Center (ASC). Cost/Schedule Data Center Brochure. Cost Division, Comptroller Directorate, ASC, WPAFB OH, 1996.

Air Force Material Command (AFMC). Contractor Performance Assessment Reporting Systems (CPARS). AFMC Instruction 64-107. 5 Aug 1997.

Brislawn, Matthew E. and Lee Dowd. "Past Performance in Government Source Selection: An Industry Perspective," Contract Management, 36: 16-18, 30 (October 1996).

Burman, Allan V. "Putting Past Performance First," Government Executive, 29: 60, (October 1997).

Chakravarthy, Balaji S. "Measuring Strategic Performance," Strategic Management Journal, 7: 437-458 (September/October 1986).

Department of the Air Force. Air Force Acquisition Fact Book. SAF/AQ Release, Washington DC, December 1996.

Department of Defense Acquisition Reform (DoDAR) Website. n. pag. http://www.acq.osd.mil/ ar/pastperf.htm. 4 July 1998.

Devore, Jay L. Probability and Statistics for Engineering and the Sciences (Third Edition). Pacific Grove CA: Brooks/Cole Publishing Company, 1991.

"DoD Sets Contractor Standards," Government Executive, 30: 10 (February 1998).

Federal Acquisition Regulation (FAR) 15.101. http://www.arnet.gov/far/html/15.html, 31 July 1998.

Gansler, Jacques S. "Collection of Past Performance Information in the Department of Defense," Memorandum from Undersecretary of Defense to the Services, Washington DC, 20 November 1997. http://www.acq.osd.mil/ar/doc/collect.htm. 4 July 1998.

Han, Sang-Lin, David T. Wilson, and Shirish P. Dant. "Buyer-Supplier Relationships Today," Industrial Marketing Management, 22: 331-338 (November 1993).

Hanson, Roger. "Revisions to CPARS Instruction/Forms." CPARS Summit Briefing, HQ AFMC/PK, WPAFB OH, 30 April 1998.

Ichniowski, Tom. "OMB Test Weighs Past Work," ENR, 232: 11 (February 1994).

Ichniowski, Tom with Debra K. Rubin. "Driven to Reform," ENR, 233: 82-84 (September 1994).

Kachigan, Sam Kash. Multivariate Statistical Analysis: A Conceptual Introduction (Second Edition). New York: Radius Press, 1991.

Larson, Paul D. and Jack D. Kulchitsky. "Single Sourcing and Supplier Certification: Performance and Relationship Implications," Industrial Marketing Management, 27: 73-81 (January 1998).

Laurent, Anne. "All Eyes on Acquisition Reform," Government Executive: 23-25 (August 1997).

Little, Arthur D., Inc. "Contractor Evaluation Program: Final Report for the Contractor Past Performance Systems Evaluation Study to the Deputy Under Secretary of Defense (Acquisition Reform)." Contract DAWS01-95-F-2182, 17 June 1996. http://www.acq.osd.mil/ar/little.htm. 4 July 1998.

Longuemare, R. Noel. "Past Performance Evaluation Rating Scale." Memorandum from Acting Undersecretary of Defense (Acquisition and Technology), Washington DC, 11 August 1997. http://www.acq.osd.mil/ar/doc/ppscale.pdf. 4 July 1998.

National Science Foundation (NSF), Division of Science Resources Studies, Federal R&D Funding by Budget Function: Fiscal Years 1996-98, Special Report, NSF 98-301, Arlington VA. 1997. http://www.nsf.gov/sbe/srs/nsf98301/pdf/histable.pdf. 4 July 1998.

Neter, John and others. Applied Linear Regression Models (Third Edition). Irwin: Chicago, 1996.

President's Blue Ribbon Commission on Defense Management. A Formula for Action, A Report to the President on Defense Acquisition. Washington DC, April 1986.

Scott, Shelley. "Best Value Contracts: Lessons Learned in Paving the Road to Quality," National Contract Management Journal, 26: 1-11 (1995).

Sledge, Avery. "Collection of Past Performance Information." Briefing, HQ AFMC/PKP, WPAFB OH, January 1998.

Spekman, Robert E. "Strategic Supplier Selection: Understanding Long-Term Buyer Relationships," Business Horizons: 75-82 (July-August 1988).

Stevens, James. Applied Multivariate Statistics for the Social Sciences (Second Edition). Lawrence Erlbaum Associates: Hillsdale NJ, 1992.

Sumpter, LeAntha. "Brief to CON 301 on DoD's Past Performance Policy." Briefing. Deputy Under Secretary of Defense, http://www.acq.osd.mil/ar/doc/con301.pdf. May 1998.

Timmerman, Ed. "An Approach to Vendor Performance Evaluation," Journal of Purchasing and Materials Management, 22: 2-8 (Winter 1996).

Weidenbaum, Murray. "What Should Defense Contractors Do?" Across The Board, 29: 50-51 (October 1992).

Wright, Jonathan L. An Examination of the Past Performance Procedures Used in Air Force Materiel Command Source Selections. MS thesis, AFIT/GCM/LAS/97S-15. School of Logistics and Acquisition Management, Air Force Institute of Technology (AU), Wright-Patterson AFB OH, September 1997 (AD-A329944).

U.S. Bureau of the Census. Statistical Abstract of the United States: *1996* (116th Edition). Washington DC, 1996.

Vita

Capt John C. Odum was born on [redacted]. He graduated from Meade County High School in Brandenburg, KY in 1989 and entered the United States Air Force Academy in Colorado Springs, Colorado. He graduated with a Bachelor of Science degree in Applied Mathematics and received his regular commission on 2 June 1993.

His first assignment was at Wright-Patterson AFB as a program manager. He has since cross-trained to become a financial manager while serving as a budget officer in the F-22 SPO. In May 1997, he entered the School of Logistics and Acquisition Management. Capt Odum has two children: John C. Odum II (Chris) and Brittney M. Odum, ages 6 and 3.

Permanent Address: [redacted]

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 1998	Master's Thesis

4. TITLE AND SUBTITLE

AN INVESTIGATION OF THE RELIABILITY OF THE CONTRACTOR PERFORMANCE ASSESSMENT REPORTING SYSTEM (CPARS)

5. FUNDING NUMBERS

6. AUTHOR(S)

John C. Odum, Captain, USAF

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
2950 P Street
WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT/GCA/LAS/98S-6

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

ASC/SYG
2335 Seventh Street, Room 2430
WPAFB OH 45433-7805

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 Words)*

An Industry trend is to establish long-term relationships with reliable suppliers. One of the criteria used to pick these "reliable suppliers" is past performance. The Department of Defense is also attempting to capitalize on this logical trend to the maximum extent possible by using past performance as an evaluation factor in source selections. Air Force Material Command (AFMC) employs the Contractor Performance Assessment Reporting System (CPARS). This thesis examines the reliability of the CPARS.

This thesis found, despite the fact that policy mandates color ratings be based on period objective measures, the cost color ratings were more consistent with cumulative objective measures. Even so, the strength of this relationship has degraded significantly over time. With respect to schedule, the reliability is improving significantly, but period objective measures are not yet significantly correlated with schedule color ratings. The author recommends that AFMC either change CPARS cost rating policy to reflect the use of cumulative objective measures or provide additional training so evaluators better understand what is assessed during a CPARS rating period.

14. Subject Terms

Reliability, Performance, Contractors, Contract Proposals

15. NUMBER OF PAGES 103

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

AFIT Control Number AFIT/GCA/LAS/98S-6

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. **Please return completed questionnaire** to: AIR FORCE INSTITUTE OF TECHNOLOGY/LAC, 2950 P STREET, WRIGHT-PATTERSON AFB OH 45433-7765. Your response is **important.** Thank you.

1. Did this research contribute to a current research project? a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?
a. Yes b. No

3. **Please estimate** what this research would have cost in terms of manpower and dollars if it had been accomplished under contract or if it had been done in-house.

Man Years___________ $___________

4. Whether or not you were able to establish an equivalent value for this research (in Question 3), what is your estimate of its significance?

a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments (Please feel free to use a separate sheet for more detailed answers and include it with this form):

____
Name and Grade

____
Organization

____
Position or Title

____
Address