JOURNAL OF PUBLIC PROCUREMENT, VOLUME 4, ISSUE 2, 210-237 2004

EVALUATION OF COMPETITIVE SOURCING PERFORMANCE WORK STATEMENT METRICS

Ty A. Randall, Heidi S. Brothers and Daniel T. Holt*

ABSTRACT. Competitive sourcing is the government's term for transferring the operation of an internal process or function to either an external supplier or a reengineered government team. The competitively sourced function is managed through performance metrics. These metrics must be thorough, appropriate and well designed to ensure the government is receiving the level of service required to fulfill its various missions. This research effort develops a performance metric evaluation system that was synthesized from metric design literature, Total Quality Management concepts, and the Government Performance Results Act. Use of the system in a case study is discussed along with how to evaluate the results. Results indicate that some Air Force performance metrics have insufficient and improperly designed metrics.

INTRODUCTION

Competitive sourcing is the federal government's term for outsourcing. Outsourcing is a strategic term used by private sector companies seeking to focus their company's skill and resources on core competencies by transferring the ownership of certain internal processes or functions to an external supplier (Outsourcing Center, 2001). By

---

* *Ty A. Randall, MS, has served the Air Force Civil Engineering organization at a number of bases in the United States and overseas. Heidi S. Brothers, PhD, is an Assistant Professor of Civil Engineering, Civil Engineering Department, United States Air Force Academy. Her teaching and research interests include civil engineering contract management, facility management, and other engineering management topics. Daniel T. Holt, PhD, is an Assistant Professor of Management, Department of Systems and Engineering Management, Air Force Institute of Technology. His research interests include organizational change, turnover, human resource management, personality, and survey development.*

outsourcing certain functions, companies can focus on their core competencies where they obtain "definable preeminence" and can offer their customers unique value. Results of outsourcing can be significant because the outsourcing company gains the utilization of an external company's resources, to include investments, innovations, and specialized capabilities, all of which would be too expensive or impossible to duplicate internally. Outsourcing can lower investments and create better responsiveness to customer needs (Quinn & Hilmer, 1994). This reasoning suggests a company can improve efficiency through the use of an external company's capabilities, such as its innovations and specialized capabilities, and improve efficiency by potentially improved customer responsiveness.

Other outsourcing research efforts have realized less than positive results. Prager and Desai (1996) evaluated an outsourced public-sector function to see if there were, in fact, savings and increased efficiencies. Their results were less than flattering and did not yield solid support for outsourcing. An analysis was conducted to determine if quality and efficiency increased as cost decreased by outsourcing foster care at group homes. Significant problems were discovered: a) the efficiency and effectiveness of processes were difficult to define, b) cost data was not maintained on in-house functions to make the post outsourcing comparisons, and c) many contracts were not monitored efficiently (Prager & Desai, 1996). In the end, the study was abandoned due to the lack of pre-outsourcing financial data. Still, the authors of the study warned against getting too involved in outsourcing without considering long term effects (e.g., higher hidden costs regardless of initial savings).

One long-term effect to be considered by leaders is the performance of the service providers that are selected. Competitive sourcing should comply with the requirements of performance-based contracting as specified in the Federal Acquisition Register Subpart 37.6. Performance-based contracts describe the requirements in terms of results required and use measurable performance standards. Statements of work in the PWS and contracting documents shall define requirements in clear, concise language identifying specific work to be accomplished "Subpart 37.6 Performance-Based," 2001). Indeed, officials have been told to incorporate clear performance metrics into the contracting process to properly evaluate service providers. Performance metrics are "a common and mutually reinforcing focus on achieving program results and customer satisfaction, measuring performance, and using

performance data to identify and select improvement opportunities" (U.S. GAO, 1999b). Performance metrics should address both financial (cost management) and non-financial (i.e., productivity, quality, timeliness, and responsiveness) activities (Buchheim, 2000). If performance metrics are written well, they will provide a method to evaluate the service provider to ensure the Government is receiving the best value for their money while meeting mission requirements (U.S. GAO, 2001).

Unfortunately, competitive sourcing actions and associated metrics have been evolving for at least 20 years and their effectiveness is not known. Yet, the efficiency improvements that organizations hope to realize with the outsourcing strategy can only be evaluated if a properly designed performance measures have been used. Therefore, the purpose of this research was to develop a method to evaluate performance metrics for competitive sourcing efforts in order to design better, more thorough metrics. Moreover, this method is applied to evaluate metrics that have been included in several competitive sourcing efforts. Our discussion will unfold by first explaining some of the background behind the Department of Defense's (DOD) outsourcing efforts. This discussion is followed by a detailed evaluation system that can be used to design better, more thorough metrics. Finally, this evaluation system is applied to several specific competitive sourcing efforts so that the metrics from these efforts can be evaluated.

BACKGROUND OF DEFENSE DEPARTMENT OUTSOURCING EFFORTS

The US Government is turning to an outsourcing strategy as a way to improve efficiency and reduce costs. The DOD's Competitive Sourcing began in 1955 and continued when the Office of Management and Budget (OMB) initiated Circular A-76 in 1966 (U.S. GAO, 1999a). Circular A-76 directed any government activity identified as a commercial activity (CA), and not classified as inherently governmental, to be competed against the private sector; government in-house bids would compete against private sector bids for the CA which is simply defined a service that is available in the private sector (USDATL, 2000).

The DOD outsourcing program, called competitive sourcing, includes a government team in the competition for the "contract." The goal of competitive sourcing is to provide services at minimum costs

while sustaining or improving performance (USDATL, 2000). Through competitive sourcing, the government focuses on core mission competencies and service requirements (OMB, 2001a). Even more specific, the four goals of the U.S. Air Force (USAF) Competitive Sourcing program are: (a) sustain readiness; (b) improve efficiency and reduce costs; (c) create funds for force modernization; (d) focus on core Air Force missions (SAF, 1997).

The entire competitive sourcing initiative is expected to trim costs by competing approximately 230,000 DOD positions between Fiscal Years 1997 and 2005. The expected savings over the same eight years is 11.2 billion dollars, which is to be reallocated for other defense priorities (DOD, 1999). Also, competitive sourcing is intended to promote competition, which motivates improved performance at reduced costs (USDATL, 2000). Therefore, it is vital in these budget-crunched years that competitive sourcing is properly initiated and evaluated.

DEVELOPMENT OF EVALUATION SYSTEM

An evaluation system was developed from metric design literature so that it would be consistent with the Total Quality Management literature (Cohen & Brand, 1993); Government Performance and Results Act (U.S. GAO, 1996); and Six Sigma literature (Wiklund & Wiklund, 2002). In addition, inputs from Air Force (AF) Civil Engineering metrics were included. Both qualitative and quantitative evaluation measures are used to determine if the performance metrics are effective in supporting the mission.

The performance standards and measures in the Performance Work Statement (PWS) and Quality Assurance Surveillance Plan (QASP) form the basis of the measurement program used to evaluate the CA performance. The performance standards are the objectives being sought by the task, and the performance measures are the metrics used to measure the results of the task. Managers use measurement programs to inform their people about the processes of the organization. Key elements within the measurement program are specific, quantifiable performance metrics; metrics are quantitative values obtained by measuring certain characteristics of a process (Edberg, 1997).

Problems have arisen when developing a metric because the metric may seek the wrong data or the metric may be confusing or not quantifiable. An example of a metric that is not quantifiable would be if

the standard stated "Operate, maintain, and repair the heating system" with a threshold of 100% of the time. The "operate, maintain, and repair" tries to include too much into the one metric and is very broad. It is not understandable to operate, maintain, and repair 100% of the time.

To help eliminate the problems with initiating a measurement program, a nine-step guideline has been established by the Department of Energy. The steps of this process are depicted in Table 1. The steps

TABLE 1
Steps to Create a Performance Measurement Program

Steps	Purpose
Involve all affected organizations in the development of performance metrics	Ensure that all affected organizations will accept the results of the effort
- Flowchart the applicable process - Determine what is important to the customer	Identify critical activities (i.e. "control points") to measure, and the results which are worthy of being measured
- Establish the performance measurements (i.e. unit of measure, sensor, and frequency) - Establish goals or standards - Identify responsible parties for data collection, analysis, and reporting	Collect the data, and ensure that the data collection process functions properly
- Analyze and report the actual performance - Compare actual performance with standard or goal - Evaluate causes of variances, and potential corrective actions	Determine what actions should be taken in response to a variance. It may be appropriate to: ➢ Ignore it (if the variance is not statistically significant) ➢ Fix it (if it is significant, or indicates an unfavorable trend) ➢ Challenge the goal (if achieving the goal would be counter-productive to more important Corporate objectives) ➢ Challenge the metric (if the metric is providing useless or hard-to-interpret information)

Source: Adapted from Buchiem (2000).

include identifying all organizations affected by the task and determining all requirements; establishing standards, metrics and a collection procedure; and reviewing the performance and identifying improvement opportunities.

The "establish the performance measurements" step (see Table 1) discusses the formation of metrics. The metric is used to evaluate the function and it must be properly designed to assess the function properly. A well-designed performance metric must include three elements: "1) a defined measure of unit, 2) a 'sensor' which gathers and records the raw data, and 3) a frequency with which measurements and reports are to be made" (Buchheim, 2000, p. 310). The performance metric must also be: (a) understandable; not difficult to define or understand, (b) quantifiable; objective with much of the personal influence or judgment reduced, (c) cost-effective; value of information sought must exceed the cost of data collection, (d) proven; validated to have shown a drive to improvement, and (e) high impact; collection of metrics must be worthwhile (Edberg, 1997).

An example of a metric that meets all these criteria might be: respond to and complete emergency, urgent and routine service calls with a required performance of 95% of the required time limits. Emergency calls must be answered with in 24 hours, urgent within five business days, and routine within one month. Information about the service calls is maintained in a database and checked monthly for compliance to the required level of performance. This metric has a defined unit of measure - time to respond, a sensor - the data of service call responses, and a frequency - once a month check. The metric is understandable and quantifiable; it is based upon the length of time to respond and complete. The metric is also cost effective because it is not labor intensive to track and is easily maintained (assuming there is an established data system). The service call metric is proven within organizations and is high impact because it shows if the organization is improving in responding to and completing the number of service calls, which also improves quality of life and sustains mission capability. Unfortunately, this metric cannot be used to evaluate quality of job completed, customer satisfaction, or budgetary performance. The developer of the PWS must decide what aspects of the function are important and decide if each of the areas must have their own metrics.

Identifying the metrics to be used in evaluating any process is not an easy task. First, it is important to recognize critical areas that need to be

evaluated. The critical areas are areas deemed important to the success of the organization and should be focused around the mission of the organization. Therefore, there needs to be some identification of critical areas to guide metric development. Recent outsourcing studies in the private sector have metrics categorized within nine gauge clusters: finance/budget, customer satisfaction, work product delivered, quality, time/schedule, business value, operational service levels, human resources, and productivity (Rubin, 1997). Table 2 lists and defines the nine clusters. The gauge clusters can help managers classify requirements into categories and then design and use metrics appropriately in the oversight of the requirements.

TABLE 2
Oversight Framework for Performance Metrics

Gauge Cluster	Objective
Finance/Budget	Cost management and on-cost delivery of services
Customer Satisfaction	Critical attributes that generate satisfaction with services and work products among internal business customers
Work Product Delivered	Quantifying the amount of service or work provided in a given time period
Quality	Objective and measurable aspects of quality of services and products
Time/Schedule	Critical service, product, and project time frames and the ability to deliver on-time
Business Value	Measures the outsourcing agreement's outcome attainment from the financial/shareholder view, external customer/marketplace view, organizational learning and improvement view, and internal process improvement view
Operational Service Levels	Critical service tempos, availability, and delivery of work products
Human Resources	Changes to the skill inventory and internal job satisfaction
Productivity	Efficiency of the production and delivery of work products

Source: Adapted from Rubin (1997).

The gauge clusters used in private sector outsourcing can be used as a starting point in the evaluation process to ensure the organization undergoing the competitive sourcing process is evaluating critical areas. The gauge clusters can be applied to the performance measures currently used in the competitive sourcing process to ensure all aspects of the functions are being assessed (i.e., financial, customer satisfaction, productivity).

The government uses many performance measurement programs. Two of the programs include the Total Quality Management (TQM) and the Government Performance and Results Act (GPRA). The Air Force endorsed performance measurement with TQM. The push for TQM in the private sector began in the 1980's and the government followed shortly after and established TQM metrics through the early 1990's. The features of TQM are focused on: customer satisfaction, employee involvement and continuous improvement (Cohen & Brand, 1993, pp. xii-xv). TQM was an initiative to improve the effectiveness and performance of the federal government (U.S. GAO, 1999b).

Establishing performance measures to evaluate the progress made by TQM initiatives is a vital process within TQM itself. The eight steps of TQM are identified in Table 3. In step seven, performance measures must be established to determine improved performance. The eight steps

TABLE 3
Total Quality Management Steps

Step	Objective
1	Identify what processes need improving; begin with identifying customers and their needs
2	Describe the steps taken in performance of the work
3	Identify the parts of the process where defects, delays, or rework occur frequently
4	Identify the causes for the defects, delays, or rework
5	Improve process by experimenting with small-scale pilot projects
6	Based on positive pilot project results, institute new procedure
7	Continually monitor the new process to ensure it improves performance over time
8	Repeat steps 1 through 7 to continually improve processes

Source: Cohen and Brand (1993, p. 6).

of TQM involve identifying key outputs and customers, identifying process areas creating delays or defects, instituting changes, and finally measuring the results and comparing them to what the customers really desired from the process.

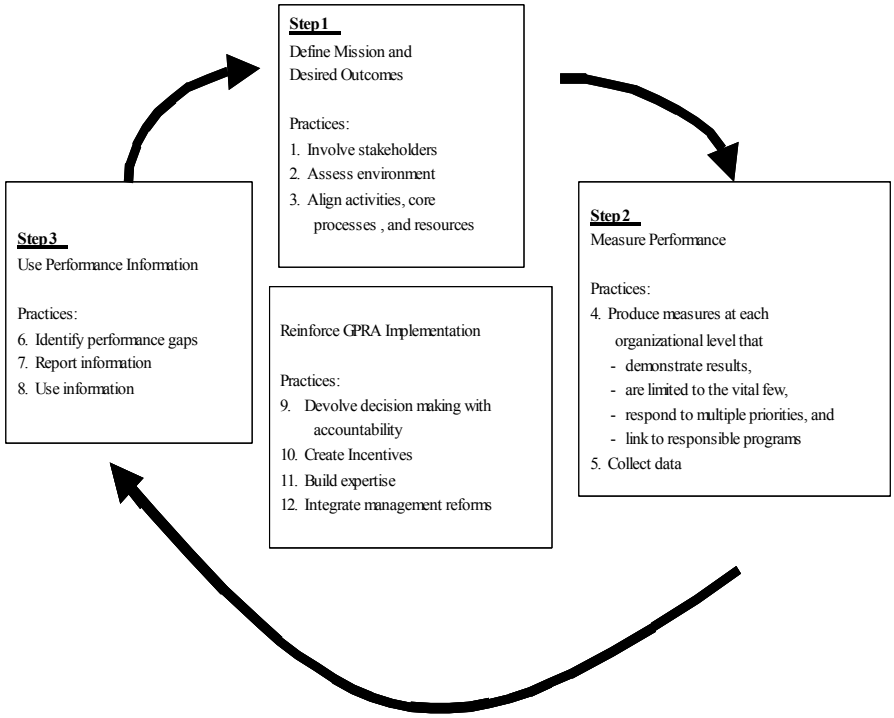
As a result of the pursuit for improved efficiency and effectiveness, and to continue the concepts of TQM, the GPRA was passed in 1993. The GPRA required Government agencies to set goals, measure the performance while seeking a goal, and report their results. Much of the goal setting and performance measurement is similar to that found in TQM. The shift resulting from the initiation of the GPRA was from staffing and activity levels to results (U.S. GAO, 1996). Figure 1 displays the concept and steps taken within the GPRA to focus on results. The three main steps involve defining the mission and outcomes desired, measurement of the performance, and utilization of the performance information to improve the process. Practices 9-12, reinforcement of GPRA principles, apply throughout the use of the GPRA.

Within GPRA, four characteristics of successful performance measures were identified. They are: demonstrate results, limit to vital few (measuring a few critical areas), respond to multiple priorities, and link to responsible programs. The characteristics are listed in Table 4 with their respective objectives.

The features of both TQM and GPRA provide an initial insight into identifying critical mission requirements, which leads to identifying key customers and outputs, and finally, building metrics to evaluate performance. The GPRA can be used further to assess the metrics; the four characteristics of the performance measures found in Table 4 along with the performance measurement data can be used to evaluate the extracted metrics. A performance metric evaluation system was established from the various metric design literatures, features from TQM, GPRA, and current civil engineering metrics were incorporated into this system.

Each metric is to be evaluated using the 11 criteria. The first six criteria contained in the evaluation step are: (a) defined measure of unit, (b) sensor, (c) frequency, (d) understandable, (e) quantifiable, and (f) high impact. Buchheim (2000) developed criteria 1, 2, and 3 and Edberg (1997)developed criteria 4, 5, and 6. The six criteria tie back to TQM

FIGURE 1
Government Performance and Results Act



Source: Adapted from U.S. General Accounting Office (1996c).

and GPRA because the criteria indicate that a measurement system is in place to evaluate the process (as specified by TQM and GPRA principles).

Criteria 7 - 11 provide a quantitative evaluation of the metrics by using specific information from the organization with the metric. The five criteria contained in the quantitative part of the evaluation are: (a) is objective measured (Criteria 7), (b) where is collected information stored (Criteria 8), (c) how long is the information stored (Criteria 9), (d) cost effective (Criteria 10), and (e) proven (Criteria 11). Criteria 7, 8, and 9 are developed from TQM and GPRA and criteria 10 and 11 are developed from Edberg (1997).

TABLE 4
Characteristics of Successful Performance Measures

Characteristic	Objective
Demonstrate Results	"Performance measures should tell each level how well it is achieving its goals" (U.S. GAO, 1996, p. 24)
Limited to vital few	"The number of measures for each goal at a given level should be limited to the vital few. Those vital few should cover the key performance dimensions that will enable an organization to assess accomplishments, make decisions, realign processes, and assign accountability" (U.S. GAO, 1996, p. 25).
Respond to multiple priorities	Performance measurements must take into account all competing interests: quality, cost, customer satisfaction, stakeholder concerns, and other factors (U.S. GAO, 1996, p. 25).
Link to responsible programs	"Performance measures should be linked directly to the offices that have responsibility for making programs work"; helps to reinforce accountability and helps managers to strive for goals (U.S. GAO, 1996, p. 25).

The evaluation system includes a "yes/no" block to identify if the evaluater felt the metric passed or failed the criteria ("no" indicates a fail). Finally, the evaluation system has a justification block to allow the evaluater to explain the given line's success or failure for each criterion.

EXAMPLE APPLICATION

The performance metric evaluation process was applied to a select set of competitive sourcing documents from USAF bases. Eight USAF bases were selected to provide examples for this research effort. The Operations Flight within the Civil Engineering Squadron was selected as a focus for this research because many Operations Flight offices have been through the competitive sourcing process. Only major competitive sourcing efforts were used (large portions of the organization). The Operations Flight is responsible for repair, maintenance and minor construction of the facilities and infrastructure of USAF bases. The eight bases were selected using four selection criteria: (a) large scope of contract, (b) major command the base supported, (c) location of base, and (d) outcome (reengineered government team vs. contract or direct

conversion). The criteria were used to identify a wide diversity of performance metrics for evaluation.

In USAF competitive sourcing actions, performance metrics are documented in the PWS and the QASP. The PWS defines what service is being requested, the measurements of performance (standards and metrics), and timeframes required. The PWS should be performance oriented by specifying what outcomes and measures are desired and not placing directions on how to achieve the outcomes and measures (OMB, 2001b). The PWS includes: a Description of Services, a Service Delivery Summary (SDS), and other contract information. The metrics used to measure the performance of the service provider are sometimes located in the QASP. The QASP is written to cover the life of the service contract and contains methods of surveillance, the performance metrics, and sometimes, incentives tied to the performance of the service provider. The SDS portion of the PWS was used to identify the performance metrics for evaluation. An example of a portion of the information found within a SDS is shown in Table 5 which was extracted from a civil engineering squadron PWS template provided on the Air Force Civil Engineering Support Agency (AFCESA) website.

Table 5 identifies the performance objective to be completed by the service provider. It identifies the "what" of a task. The table also

TABLE 5
Service Delivery Summary Example

Performance Objective	SOW Paragraph	Performance Threshold
Treat customers politely, cheerfully and promptly	1.1.1	Customer service rating of at least 4.0 on a 5.0 scale
Respond to and complete emergency, urgent, and routine service calls	1.1.3.	95% of required time limits
Maintain, repair, construct, and operate the supporting infrastructure ensuring cost effective and reliable support	1.2.	100% of time
Provide economical maintenance, repair, construction, installation, operation, and service functions for real property, Real Property Installed Equipment (RPIE), and designated Equipment Authorized Inventory Data (EAID)	1.3.	95% of scheduled inspections and/or work completed on time

identifies the paragraph of the Statement of Work (SOW) where the requirement can be found. Each requirement has a specific location within the SOW or PWS and is typically identified by chapter numbers, followed by paragraph and sub-paragraph numbers. For example, Table 5 contains a performance objective stating the service provider must "treat customers politely, cheerfully, and promptly", which can be found within sub-paragraph one, of the first paragraph of Chapter One of the particular SOW. Finally, Table 5 establishes the threshold value to which the service provider will be held accountable. The information contained in the performance threshold should not dictate to the service provider "how", but establish the threshold of performance that will be measured (the results of the task). The SDS portions of the PWS documents provide the minimum performance required of the service provider.

Competitive sourcing documents are not consistent across the USAF. Sometimes performance metric information appears in the QASP. The QASP is written to assist the contracting officer's performance evaluation of the service provider and defines the process by which the Government will evaluate the performance of the service provider and evaluate the compliance of the service provider with PWS standards (USDATL, 2000). Sometimes the QASP contains the frequency, purpose, and method of each inspection, along with the penalties of not meeting the performance standards listed within the PWS.

Performance metrics must be directly linked to the mission requirements of the organization. There is no value in measuring performance of an activity that does not support the mission. For this research effort, the mission requirements of the Operations Flight office were identified. The primary responsibilities of the Operations Flight office are to "ensure Air Force installations can support the mission, maintain real property facilities, and develop and implement programs to improve the livability of our base community" (DAF, 1999). The Operations Flight is tasked with fourteen functions, identified from Air Force Instruction (AFI) 32-1001 and listed in Table 6**.** Table 6 lists the function identification as it appears in AFI 32-1001, the function itself, and then a brief reference name to be used for the remainder of the research.

Evaluating existing metrics requires several steps. First, the metric must be dissected into objective, threshold and surveillance. An example

TABLE 6
Civil Engineer Operations Flight Functions

Function ID	Function Objective	Reference Name
1.1	Operate, maintain, repair, construct, and demolish Air Force real property and real property installed equipment (RPIE) to accomplish the mission in the most timely and economical manner	Operate, maintain, and repair
1.2	Provide trained personnel and technical expertise to support Air Force operations worldwide	Trained personnel
1.3	Maintain capability to respond to and eliminate any emergency condition 24 hours a day	Emergency response
1.4	Conduct all activities in compliance with applicable environmental, fire and safety laws, codes, and directives	Compliance
1.5	Provide reliable, cost-effective utilities to meet readiness requirements, satisfy installation needs, and maintain quality of life	Reliable utilities
1.6	Provide base support services (i.e., pest control, grounds maintenance, snow removal)	Base support
1.7	Establish quality standards and feedback mechanisms to assess performance in meeting mission requirements and customers' needs	Quality standards
1.8	Establish a system to provide customers the capability to accomplish work requirements using their own resources	Self help
1.9	Develop and annually update future plans for major work requirements (roofing, pavements, protective coating)	Future plans
1.10	Effectively allocate in-service resources, including people, facilities, equipment, and vehicles to meet mission and customers' needs	Allocate Resources
1.11	Provide customers with the costs of work or services performed on their facilities	Provide costs to customers
1.12	Maintain a time and material accounting system to collect and report the cost of doing business	Time and material accounting
1.13	Provide effective logistics support	Logistics support
1.14	Provide an effective facility manager program	Facility manager program

Source: Department of Air Force (1999).

of such a dissection of a metric from an AF civil engineering PWS is shown in the first four columns of Table 7**.** The first column is an identifying number. For this research, the metric dissection and placement into the table is referred to as a line. The creation of the lines allows for easier evaluation of the metrics and a line is created and numbered for each metric to be evaluated.

Next, each line is matched with the appropriate mission requirement from the organization documents. Table 7 provides an example identifying the corresponding mission requirement (in this case one of the 14 civil engineer operations flight functions from Table 6) in the fifth column labeled Operations Flight Function. The classifications of the lines to appropriate mission requirements indicate what areas of the organization are evaluated and to what extent the areas are evaluated (some areas may have more than one metric measuring the area).

TABLE 7
Example Metric Line

ID	Objective	Threshold	Surveillance	Operations Flight Function	Gauge Cluster
1	Provide production control that is professional and courteous at all times.	0 Defects. Lot is number of calls received or verbal requests taken monthly.	Customer Complaint	1.2, Provides trained personnel and technical expertise to support operations worldwide	Customer Satisfaction

Next, the metric line is classified into one of the nine gauge clusters from Table 2. This is shown in the last column of Table 7. The classifications of the lines into gauge clusters determine what aspect of the mission functions is evaluated (e.g., quality, customer satisfaction, or cost). The gauge cluster classification was based upon information contained in the threshold and the objective of the line.

If an objective and threshold were found to contain two or more gauge clusters, (for example time/schedule and quality) then the objective and threshold were broken into two separate lines, one for the

time/schedule gauge cluster and another for the quality gauge cluster. This was done to reduce the confusion of how to classify the line.

Line Evaluation

After the metrics are identified in the line format, they are evaluated individually. For each metric line, an evaluation table with the 11 criteria questions is completed. In the case study, 161 metric lines were identified and evaluated.

The metric line evaluation was broken into two parts for the research. The first part included criteria 1-6 and the second part included criteria 7-11. Only those metrics passing the first six criteria were subjected to the secondary evaluation because the secondary part required additional information from the original organization. The purpose of the secondary evaluation was to provide a quantitative evaluation of the metric lines that passed the primary evaluation. An example of a metric line evaluation using criteria 1-6 is shown in Table 8 and 9.

TABLE 8
Example of a Line without a Defined Measure of Unit

ID	Objective	Threshold	Surveillance
160	The service provider will be expected to maintain a professional appearance of all sites, roads, airfield pavements, parking areas, etc. Any deviation from this professional level of service (e.g., complaints of pot holes, low areas holding water on the flightline, spalls on the airfield, accumulation of debris on streets/pavements, higher counts than threshold limits for mosquito population, etc.) will be counted as a data point in this metric. An incident is the first time that an event is brought to the attention of the service provider. A complaint is dissatisfaction with the fix or a lack of response to the incident. This metric will have two columns; one for an incident and one for complaints.	No deviations	Monthly

TABLE 9
Evaluation of Line #160

Criteria	Yes/No	Justification
Defined unit of measure?	No	There is no specified unit of measure
Sensor?	Yes	Records
Frequency?	Yes	Monthly
Understandable? (Not difficult to understand)	No	Due to lack of specified unit of measure
Quantifiable? (Reduced personal influence or judgment)	No	Professional appearance cannot be quantified
High Impact? (Affect Quality of Life, Mission, or Customer)	Yes	Has significant impact upon the customers of the civil engineering, quality of life, and successful completion of the mission

Of the 161 metrics evaluated, only 33 passed the initial evaluation using criteria 1-6. And of those, only seven metrics could be evaluated with criteria 7-11 because the other 26 came from bases that were in the process of awarding the competitive sourcing contracts (and therefore did not have the quantitative information needed for the second part of the evaluation). Tables 8 and 9 contain an example of a line that failed the evaluation of the first six criteria. Note the three "no" answers in the second column of Table 9. The mission function and the gauge cluster classifications were removed from Table 8 because they are not needed during this part of the evaluation. There is no defined unit of measure for the threshold, and the surveillance. The “maintain a professional appearance” can be found in the objective, but professional appearance is not something that has a defined unit of measure (i.e., time, cost).

Tables 10 and 11 provide an example of the evaluation of criteria 7-11. The information used to complete Table 11 was obtained from phone interviews with the organization.

The result of the metric evaluation identifies metrics that are properly (or improperly) designed. Those metrics failing some portion of the evaluation can be reviewed to determine recommendations for improvement. After all metric lines are evaluated, there are several ways to interpret/use the information. For the case study, a table was prepared that compared the number of metrics corresponding to each mission

TABLE 10
Example Line Passing the Secondary Evaluation

ID	Objective	Threshold	Surveillance
5	Emergency Work Requests: 30 minutes (duty hours)/1 hour (non-duty hours); completed (safed) in 24 hours.	100% of the time	Records Review or Customer Contact, at least 1/week

TABLE 11
Secondary Evaluation of Line #5

Criteria	Yes/No	Justification
Is objective measured?	Yes	By contractors telling QAE he will not make the required time
Where is collected information stored? (Accessible to those that need information)		IWIMS (computerized data base)
How long is the collected information stored?		1 year
Cost Effective? (Value of obtained information outweighs cost of seeking information)	Yes	Very little cost involved
Proven? (Has shown demonstrated results)	Yes	If the work request is not completed in specified time, the QAE questions the service provider and initiates closure as close to the specified time as possible

function to the appropriate gauge cluster. Table 12 summarizes these numbers. Only a third of the mission functions are shown as an example. There are nine gauge clusters, but only seven were found during the classification of the 161 lines.

Table 12 shows that the majority of the metrics are concentrated on just a few mission functions. Two-thirds of the metrics are concentrated

TABLE 12
Number of Metrics per Mission Function and Gauge Cluster

Mission Functions	Gauge Clusters							
	Time/Schedule	Operational Service Level	Quality	Finance/ Budget	Customer Satisfaction	Work Product Delivered	Productivity	Total number of metrics per mission function:
1.1, Operate, maintain, and repair	14	15	6	3	0	0	0	**38**
1.2, Trained personnel	3	19	4	1	1	1	0	**29**
1.3, Emergency response	5	0	0	0	0	0	0	**5**
1.4, Compliance	0	0	0	0	0	1	0	**1**
1.5, Reliable utilities	2	31	2	0	1	0	3	**39**
7-14 functions	11	25	7	4	0	1	1	**49**
Total number of metrics found within gauge cluster classification	**35**	**90**	**19**	**8**	**2**	**3**	**4**	**161**

on mission functions 1.1, 1.2 and 1.5. If all mission functions are of equal importance, then the metrics need to more evenly distributed across the functions. A similar imbalance is indicated by the "Total number of metrics found within the gauge cluster classification" on the bottom of the table. That is, 56% of the metrics address the operational service level whereas other gauge cluster areas have only a few metrics. Very few of the lines contained a gauge classification of quality, finance/budget, productivity, work product delivered, customer satisfaction, and business value, and none of the lines contained a gauge classification of human resources. Quality and Customer Service should be integral to the evaluation of the civil engineering processes because civil engineering is a service, but quality and customer service are very difficult to quantify. It is the researchers' opinion that the imbalances are

not due to significant imbalances in the importance of the mission functions and gauge clusters. Instead, the imbalances are due to a history of metrics that only address certain areas of mission requirements (often these are easier to create and evaluate than other types of metrics.) The table shows a need for a wider diversity of metrics.

DEVELOPING BETTER METRICS

The performance metric evaluation process provides a means to identify where metrics are needed and a structure for improving existing metrics with design flaws. The evaluator can use information and ideas from all the evaluated metrics, supporting literature and other organizational documents to redesign the performance metrics.

In the case study, 19 new metrics were designed to address mission areas that did not have associated metrics and improve existing metrics that failed the evaluation. An example of a new metric from the case study that meets the evaluation criteria is shown in Table 13. The other 18 metrics are listed in Appendix A. The metric includes an appropriate objective, threshold and surveillance, which focus on the quality of the output from the system.

TABLE 13
Proposed Operate Systems Quality Metric

Systems output commensurate with industry standards	95% of all systems (as determined by leadership) measured weekly must have desired output (or a frequency established by leadership)	Review of records and customer complaints on a weekly basis (must match threshold time)

CONCLUSIONS

Competitive sourcing and outsourcing efforts are commonly used by governments and business to improve efficiencies and reduce costs. In order to realize the benefits of this competitive sourcing strategy, there must be a relevant, thorough, quality evaluation program in place to ensure mission requirements are being meet. That evaluation program is based on performance metrics.

Some literature is available on designing performance metrics including the eight steps of Total Quality Management and the concepts of the Government Performance Results Act. In this paper, a performance evaluation system was developed from the literature that uses 11 questions as a mechanism to evaluate individual metrics. This performance metric evaluation system is applicable to all competitive sourcing/outsourcing efforts and will help organizations effectively evaluate if they are getting their money's worth out of their competitive sourcing efforts.

To demonstrate this, an extensive case study was conducted using performance metrics from eight diverse USAF bases where 161 different performance metrics were evaluated. The case study results were used to demonstrate how the evaluation process can be used to identify mission requirements, areas that are insufficiently covered with metrics, and an example was provided on how poor metrics can be redesigned to be more effective.

Unfortunately, the results from the case study were unsettling. It demonstrated the current methods of measurement make it difficult for the USAF to determine whether it is getting the services required. Nor can the performance metrics be used effectively to evaluate efficiency, effectiveness or quality. Of the 161 metrics evaluated, the vast majority covered only three of the 14 mission functions and only one of the nine gauge clusters. The other mission functions areas and gauge clusters should be equally covered by quality performance metrics. As the evaluation proceeded, only 33 metrics passed the initial evaluation using criteria 1-6. And of those, only seven metrics could be evaluated with criteria 7-11 because the other 26 came from bases that were in the process of awarding the competitive sourcing contracts (and therefore did not have the quantitative information needed for the second part of the evaluation.) This presents a bleak picture of the thoroughness and quality of the competitive sourcing performance metrics used by the USAF and potentially DOD. Considerable effort should be made to implement quality, thorough metrics, like the ones developed in the case study, in all competitive sourcing efforts.

Because of the extensive efforts in competitive sourcing that are complete and underway, USAF bases would benefit from quality examples to guide them in the proper evaluation of the effort that is being competitively sourced. The examples and templates currently available on the Air Force Civil Engineering Support Agency (AFCESA)

webiste are poorly designed to properly evaluate the competitive sourcing efforts (AFCESA, 2001). The only way to determine if competitive sourcing is meeting the USAF Outsourcing and Privatization objectives of sustain readiness, improve activities, generate savings for modernization and focus on core activities is to have properly designed metrics evaluating the process. The focus of the designed metrics should consider both financial and non-financial aspects of the process. If quality metrics are not used, then the efficiency of competitive sourcing efforts cannot be determined, resulting in wasted money, personnel time and potential mission impact.

The main limitations of this study revolved around the method that was used. It was a case study. And, the performance metrics that were evaluated were from USAF civil engineering operations flight competitive sourcing efforts—a limited sample. Indeed, a wider array of metrics could be evaluated and redesigned. Additionally, the metric evaluation design and application is partially subjective and depends upon the knowledge of the user.

Moreover, this case study focused on the administration of competitive sourcing contracts, highlighting the need for effective metrics to evaluate service providers. Prior to this contract administration phase, however, leaders are faced with the fundamental decision to competitively source a function and subsequently decide whether an external (i.e., private sector agency) or internal (i.e., reengineered government agency) service organization be selected to provide that service. We recommend that researchers further explore the decision-making process and models that guide the selection of service providers. However, as we note above, we strongly encourage that these models use multi-attribute utility functions that incorporate both financial and non-financial criteria so that dimensions of quality, timeliness, and responsiveness are included.

As these non-financial criteria are considered, the appropriate focus of the metrics that are derived deserves further exploration. The gauge clusters gives some insight as to what issues should be addressed with the measures of performance. However, there may be an optimal mix of metrics that are used to ensure that critical dimensions are measured, enabling accurate performance measurement and encouraging continual improvement, while avoiding an unmanageable system that includes too many metrics. For instance, it might be appropriate to have a greater proportion of performance measures focused on timeliness (i.e.,

time/schedule gauge cluster) and quality rather than customer satisfaction because satisfaction would be an expected by-product of timely and quality service. In contrast, a satisfied customer may be the only necessary indicator that quality and timely services are provided.

The process of metric development does not, however, end with the decision on what should be measured. As we have suggested, there is a need to clearly define what is actually measured (e.g., number of untimely responses, perceptions of the service provider's friendliness). As this is done, leaders should be aware that the precision to which certain concepts can be measured varies dramatically. For instance, a standard that taps the mean time to repair should be more objectively and accurately captured when compared the more subjective standard of customer satisfaction. Thus, the developers of metrics must ensure that relevant results are measured as validly and reliably as possible, reflecting the desired performance.

In summary, our findings reinforce the need to have quality, thorough metrics in the PWS and have provided recommendations to improve the quality of existing metrics. At a more broad level, we hope to have encouraged government leaders to completely analyze the metrics used to gauge the performance of their competitively sourced functions. The recommendations made in this in this paper can be incorporated into the development of the PWS and performance based contracts. In the end, both should be developed in sufficient detail to ensure the basic requirements of the contract are conveyed along with a set of measurable performance standards that can be used to evaluate service providers.

ACKNOWLEDGEMENTS

This research was sponsored by the Chief of Competitive Sourcing and Privatization, HQ USAF, and supported by AFCESA and other organizations across the AF.

REFERENCES

Air Force Civil Engineer Support Agency (AFCESA) (2001). *Performance Metrics*. [On-line]. Available at http://www.afcesa.af.mil/Directorate/CEO/Management/Operations/metrics.doc.

Buchheim, R. K. (2000). "Developing Performance Metrics for a Design Engineering Department." *IEEE Transactions on Engineering and Management, 47* (3): 309-320.

Cohen, S., & Brand, R. (1993). *Total Quality Management in Government*. San Francisco: Jossey-Bass Publishers.

Department of the Air Force (DAF) (1999). *Operations Management*. (Air Force Instruction 32-1001). Washington, DC: HQ USAF.

Department of Defense (DOD). (1999). *Competitive Sourcing: A-76 Process*. [On-line]. Available at http://www.defenselink.mil/dodreform/1999update/4_1.html. (Retrieved June 13, 2001).

Edberg, D. T. (1997). "Creating a Balanced IS Measurement Program." *Information Systems Management*, *14*, (2): 32-40.

Office of Management and Budget (OMB) (2001a). *A-76 Supplemental Introduction*. [On-line]. Available at http://www.whitehouse.gov/omb/circulars/a076/a076spre.html (Retrieved June 13, 2001).

Office of Management and Budget (OMB). (2001b). *Testimony of John A. Koskinen Regarding the Freedom From Government Competition Act, Testimony: 24 September 1996*. [On-line]. Available at http://clinton4.nara.gov/OMB/legislative/testimony. (Retrieved August 21, 2001).

Prager, J., & Desai, S. (1996). "Privatizing Local Government Operations: Lessons From Federal Contracting Out Methodology." *Public Productivity & Management Review, 20*, 185-201.

Quinn, J., & Hilmer, F. (1994). "Strategic Outsourcing." *Sloan Management Review*, *36:* 43-55.

Rubin, H.A. (1997). "Using Metrics for Outsourcing Oversight*." Information Systems Management*, *14*: 7-14.

Secretary of the Air Force (SAF) (1997). *Outsourcing and Privatization*. (Air Force Policy Directive 38-6). Washington, DC: HQ USAF.

"Subpart 37.6 Performance-Based Contracting." (2001). *Federal Acquisition Regulations*. [On-line]. Available at www.fai.gov.

Under Secretary of Defense for Acquisitions, Technology & Logistics (USDATL) (2000). *A-76 Cost Comparison Handbook #4: Developing the Performance Work Statement (PWS) and the Quality*

*Assurance Surveillance Plan (QASP) (*Draft 1). Washington, DC: Author.

United States General Accounting Office (U.S. GAO). (1996). *Government Performance and Results Act* (GAO/GGD-96-118). Washington, DC: Author.

United States General Accounting Office (U.S. GAO). (1999a). *DOD Competitive Sourcing: Results of Recent Competitions* (GAO/NSAID-99-44). Washington, DC: Author.

United States General Accounting Office (U.S. GAO). (1999b). *Management Reform: Using the Results Act and Quality Management to Improve Federal Performance.* (GAO/T-GGD-99-151). Washington, DC: Author.

United States General Accounting Office (U.S. GAO). (2001). *DOD Competitive Sourcing: Results of A-76 Studies Over the Past 5 Years*. (GAO-01-20). Washington, DC: Author.

Wiklund, H., & Wiklund, P. (2002). "Widening the Six Sigma Concept: An Approach to Improve Organizational Learning." *Total Quality Management, 13:* 233-239.

APPENDIX A
Nineteen new performance metrics for AF civil engineer operations flight competitive sourcing effort

Objective	Threshold	Surveillance
Systems output commensurate with industry standards	95% of all systems (as determined by leadership) measured weekly must have desired output (or a frequency established by leadership)	Review of records and customer complaints on a weekly basis (must match threshold time)
Operate system	System mission impacting failure time will not exceed 5.3 minutes (99.999%) annually (or other established time)	Monthly review of records and customer complaints (time must match threshold time)

Objective	Threshold	Surveillance
Emergency (substitute urgent or routine) work requests completed in 24 hours (substitute 5 days and 30 days for urgent and routine, respectively)	95% of all systems (as determined by leadership) measured weekly must have desired maintenance (or a frequency established by leadership)	Review of records and customer complaints on a weekly basis (must match threshold time)
Maintain system to standard commensurate with design criteria and accepted industry standards	95% of all systems (as determined by leadership) measured weekly must have desired maintenance (or a frequency established by leadership)	Random sampling, review of records (parts consumed) and customer complaints on a weekly basis (must match threshold time)
Measure the cost of maintaining system	Determine some annual percentage of the total cost of the system that the maintenance cannot exceed (identifies aging systems)	Annual review of maintenance records and repair actions (time must match threshold time)
Maintain system according to schedule	95% of all systems (as determined by leadership) measured weekly must have desired maintenance (or a frequency established by leadership)	Monthly review of maintenance records and repair actions (time must match threshold time)
Emergency (substitute urgent or routine) work requests completed in 24 hours (substitute 5 days and 30 days for urgent and routine, respectively)	95% of all systems (as determined by leadership) measured weekly must have desired response timet (or a frequency established by leadership)	Review of records and customer complaints on a weekly basis (must match threshold time)
Repair system to standard commensurate with design criteria and accepted industry standards	95% of all systems (as determined by leadership) measured weekly must be repaired according to standards (or a frequency established by leadership)	Random sampling, review of records (parts consumed) and customer complaints on a weekly basis (must match threshold time)

Objective	Threshold	Surveillance
Measure the cost of repairing system	Determine some annual percentage of the total cost of the system that the repairs cannot exceed (identifies aging systems)	Annual review of repair records and repair actions (time must match threshold time)
Repair system according to schedule	95% of all systems (as determined by leadership) measured weekly must be repaired according to standards (or a frequency established by leadership)	Monthly review of maintenance records and repair actions (time must match threshold time)
Measure the number of base support commitments (number of commitments completed on time divided by the number of commitments)	95% of all support commitments (as determined by leadership) measured monthly must be completed on time (or a frequency established by leadership)	Review of records and customer complaints on a monthly basis (must match threshold time)
Installation support commensurate with industry practices	95% of all support (as determined by leadership) measured monthly must be completed according to standards (or a frequency established by leadership)	Random sampling, review of records and customer complaints on a monthly basis (must match threshold time)
Measure the number of commitments (number of commitments completed divided by number of total commitments)	95% of all support commitments (as determined by leadership) measured monthly must be completed (or a frequency established by leadership)	Monthly review of records (time must match threshold time)
Measure the number of self-help commitments completed on time (number of commitments completed on time divided by the number of commitments)	95% of self-help commitments (as determined by leadership) measured monthly must be delivered on time (or a frequency established by leadership)	Review of records and customer complaints on a monthly basis (must match threshold time)

Objective	Threshold	Surveillance
Track variance in cost estimates. Conduct analysis of all estimates having a 10% or greater difference between planned and actual cost)	95% of self-help commitments(as determined by leadership) measured monthly must not be greater than 10% variance (or a frequency established by leadership)	Random sampling, review of records and customer complaints on a monthly basis (must match threshold time)
Measure the number of self-help customers (number of self-help estimates completed divided by number of total self-help estimates)	95% of all self-help commitments (as determined by leadership) measured monthly must be completed (or a frequency established by leadership)	Monthly review of records and customer complaints (time must match threshold time)
Measure the number of plans and updates (number of plans and updates completed on time divided by the number of plans and updates)	95% of all plans and updates (as determined by leadership) measured monthly must be completed on time (or a frequency established by leadership)	Review of records and customer complaints on a monthly basis (must match threshold time)
Track variance in estimates. Conduct analysis of all estimates having a 10% or greater difference between planned and actual cost) or number of re-writes	95% of all plans and updates (as determined by leadership) measured monthly cannot exceed 10% variance between planned and actual cost (or a frequency established by leadership)	Random sampling, review of records and customer complaints on a monthly basis (must match threshold time)
Measure the number of plans and updates (number of plans and updates completed divided by number of total plans and updates)	95% of all plans and updates (as determined by leadership) measured monthly must be completed (or a frequency established by leadership)	Monthly review of records and customer complaints (time must match threshold time)