

# **PROTEST PROOF SOURCE SELECTION**

**A critical review of current assessment methods and a prescription for an improved source selection process.**

**By Expert Choice, Inc. & Battelle Memorial Institute  
October 1, 1998**

## TABLE OF CONTENTS

<b>Introduction .....</b>	<b>3</b>
<b>Problems With Using Numbers .....</b>	<b>5</b>
<b>Examples of Assessment Breakdowns .....</b>	<b>9</b>
<b>Linking Assessment to Evaluation.....</b>	<b>13</b>
<b>Source Selection Case Study .....</b>	<b>24</b>
<b>Conclusions .....</b>	<b>26</b>

## INTRODUCTION

Contracting executives in industry and government are focusing more and more on acquisition reform. The goal of reform is to make the procurement process faster, more effective, and easier. In response to this initiative, over the past three years our well-trained and experienced facilitators have developed and refined our team-based methodology for source selection. Our approach helps all concerned deal with the most painful parts of source selection like schedule delays and messy evaluations that result in lost protests.

“One of the most common, justifiable contractor complaints is that the notification of award to another contractor and the debriefing were not timely. Debriefings, they add, are sterile and don't give reasons why the award was made to someone else... This is a major cause of bid protest, because the unsuccessful offerors feel the agency can't or won't tell them why they did not win the award, maybe GAO or GSBCA will. The lack of information may imply those who run the government's evaluation and award process have something to hide.”<sup>1</sup>

“Information provided in a debriefing used to be limited to avoid providing new information that could be used in a protest. Current thinking is that direct and open feedback may reduce the likelihood of a protest and, more importantly, allows our industry partners to improve future proposals. Some contractors have protested government awards primarily to obtain a clear understanding of why they lost.”<sup>2</sup>

Suppliers need to know that the selection process is fair and logical. They want feedback on their performance in a timely fashion so that they can use the information to increase their competitiveness. They must understand how their performance was measured, so explaining a contract award or selection is a critical step in the selection process. Because assessment is the foundation of the selection process, this paper will focus on the link between evaluator judgments and high quality assessment data. High quality assessment methods, when designed into the evaluation model at the outset, make the selection process logical and the selection itself easy to explain.

Given a set of requirements, bidders submit proposals that provide purchasers with both hard data and soft data on their capabilities. Soft data has to do with intangibles and uncertainties such as past performance, reputation, future capabilities, etc. In selection processes, intangible factors are usually the most important, but they are sometimes slighted because we are trained to believe that it is not possible to measure them. We will show that this is not true and that it is, indeed, possible to create a scale for measuring and comparing intangibles and that such a scale is easy to construct and use.

Hard data might include overall cost figures or performance metrics. In some cases, it is appropriate to value hard data in a proportional manner so that, for example, a vehicle with a 30 MPG rating which is twice as fuel efficient as one with a 15 MPG rating is also twice as valuable. However, a vehicle capable of twice the maximum speed of another is not necessarily twice as valuable to you, given that each vehicle exceeds the minimum speed requirement. The important concept to remember about hard data is that it must be interpreted on a value scale, just like soft data.

The evaluation of performance against requirements is done by humans, and we recognize that humans are inherently subjective and inconsistent. Therefore, an effective process is one that elicits the best possible subjective judgments from evaluators. The Federal Acquisition Regulations underscore the importance of subjectivity:

“...The role of each member of the Acquisition Team is to exercise personal initiative and sound business judgment in providing the best value product or service to meet the customer’s needs.”<sup>3</sup>

Still, evaluators are human, and they apply any assessment scale they are given differently. For example, some grade high, and some grade low. Sometimes their judgments are inconsistent from factor to factor or bidder to bidder. In the end, individual judgments must be aggregated to form an overall evaluation team score. As a result, a successful assessment or measurement system for source selection must take subjectivity and inconsistency into account and aggregate individual, subjective judgments effectively.

Finally, we recognize that we must be prepared to simulate conditions or priorities that would alter our selection in order to demonstrate the strength of our rationale. The better we understand the dynamics of our selection model – the interaction between factors, their weights or importance, and alternative scores – the better we are able to explain and defend our selections. However, the confidence we have in our ability to defend a selection is just one side of the coin. The other side is contractor understanding, acceptance, and learning.

Our success has provided a breath of fresh air to the contracting community, and this paper is designed to communicate and share the benefits of our approach with you. While it is ideal to use our methods from the outset of the source selection process, we have found that we can ‘plug into’ or intercept an existing procurement process at almost any point and add significant value.

## PROBLEMS WITH USING NUMBERS

Choosing the winning contractor, selecting the best product, or picking a new supplier involves an assessment of how well each alternative contractor, product, or supplier satisfies the objectives or criteria being considered. The level to which an alternative performs with respect to an objective or criterion is associated with a rating scale value which might be an adjective, such as:

- Outstanding, Excellent, Good, Marginal, or Unsatisfactory
- Blue, Green, Yellow, or Red

Or numeric, such as:

- 0-10 in increments of 1
- 0-100 in increments of 5, 10, or 20
- 0-1,000 in increments of 50 or 100

At times, these approaches – adjective and numeric – are combined by associating numeric values with the adjectives, i.e., Blue=4, Green=3, Yellow=2, Red=1.

All of these methods appear to be straightforward and are quite commonly used within and outside of source selection processes, such as in performance evaluation, site selection, and funds allocation. The extent to which procurement policy is documented and codified based on methods such as these gives them the appearance of being systematic, logical, and rational. Their widespread use, however, does not make them immune from being misleading, inappropriate, or even wrong. In fact, improper use of number scales is a sound basis for many procurement protests.

Regardless of whether number scales are implicit (adjective methods) or explicit (numeric methods), not all numbers are created equal. There are actually four distinct types of number scales<sup>4</sup>:

- Nominal
- Ordinal
- Interval
- Ratio

### Nominal Scales

Numbers on a nominal scale are simply names. Phone numbers and Social Security numbers are a means of identification. The Interstate Highway numbering system uses numbers as names to designate highways. I95, I395, and I70 are examples. By convention, odd numbers indicate north-south roads, and even numbers indicate east-west roads.

No arithmetic of any kind can be performed on nominal data. No reasonable person would try to add I95 to I395, and no one would expect that the resulting ‘addition’ would run from east to west!  $1 + 1$  simply does not equal 2 (or anything else) on a nominal scale. Any source selection whose multicriteria assessment is based on nominal scales is subject to protest.

## Ordinal Scales

Numbers on an ordinal scale indicate order only. They are ranks. A political candidate or a sports team might be in 1<sup>st</sup> place, but their rank of '1' contains no inherent meaning regarding the distance from the next place candidate or team. The interval between 1st and 2nd places, for example, is not captured in the rankings themselves since the rankings only indicate order of preference, not strength of preference.

As a result, using a 1-4 (or 1-n) rating scale and multiplying these numbers by criteria weights gives meaningless results. Using such a scale imposes a forced decreasing marginal utility between increments of the scale. Evaluators usually think that the same interval has the same meaning anywhere on the scale, but this is not so. The difference between a '1' and a '2' appears to be the same as the difference between a '3' and a '4'. However, when the rank order values are mistakenly used as weights, a '2' is 100% better than a '1', a '3' is 50% better than a '2', and a '4' is only 25% better than a '3'.

No arithmetic of any kind can be performed on ordinal data, so  $1+1$  does not equal 2 on an ordinal scale, either. Any source selection whose multicriteria assessment is based on ordinal scales is subject to protest.

## Interval Scales

On an interval scale, the intervals between points are meaningful, and corresponding intervals on different parts of the scale have the same meaning. The Fahrenheit temperature scale is an interval scale. 90° F is 10 degrees more than 80°, and 45° is also 10 degrees more than 35° F, but 90° F is not twice as hot as 45°. Here's why: temperature can be measured on another scale, the Celsius scale. Using this interval scale we find that the Celsius equivalent of 90° F is 32.2, and 45° F is 7.2 Celsius, so one reading is 4.5 times the other. Because of the scale we pick, we could draw different conclusions from the same data! You can add and subtract numbers on interval scales, but you cannot multiply or divide numbers on them.

The use of interval scales in source selection involves both theoretical and practical difficulties. Theoretical problems arise whenever we subdivide evaluation factors into subfactors and subfactors into sub-subfactors and then multiply the weights from level to level. When we roll up interval scale assessments at the subfactor level by multiplying assessments against interval scale factor weights, we generate meaningless data. Any source selection involving a hierarchy of criteria whose multicriteria assessment is based on interval scales is subject to protest. Furthermore, a protest opportunity can be created even when we avoid subdividing factors. If we use a simple, two-dimensional evaluation matrix (with bidders as rows and factors as columns), the overall results (arrived at by multiplying scores times weights) are meaningless if the factor weights are determined using an interval scale.

At a practical level, when we use 1-10 scales to rate alternatives in a source selection, we leave ourselves open to the criticism that we have inaccurately measured performance against requirements. A rating of '8' (on a scale from 1-10) for one alternative and a '4' for another alternative would tempt us to conclude that one alternative is twice as preferable as the other. However, this inference is not valid unless we can be sure that each evaluator used the scale with this conclusion in mind, i.e., that by using '8' and '4', the evaluator meant to indicate that one was twice as good as the other. Consistent application of interval scales is, in practice, very difficult to accomplish.

The same difficulty arises with factor weights. If we avoid subdividing factors into subfactors, then we are forced to determine weights across a potentially large number of factors. When we give one factor an '8' and another a '4', we must mean that one is twice as important as the other to make the results valid. But what if there are 50 factors? When we reach the 50th factor, can we really expect to remember the weights of earlier factors? If the first factor got a '10' and the 50th factor a '1', do we really mean that the 50th factor is one tenth as important as the 1st?

## Ratio Scales

On a ratio scale, intervals between points on the scale are consistent and meaningful, but the zero point on a ratio scale indicates the absence of the attribute being measured. Age, distance, and money are a few of the many examples of natural ratio scales. The Kelvin temperature scale, unlike the Fahrenheit and Celsius scales, is a ratio scale. Its '0' (known as Absolute Zero) is not arbitrary since it is the point at which no molecular activity exists.

The most important characteristic about ratio scales is that all arithmetic operations on ratio data are valid. If one evaluator gives an option an '8' and another option a '4', that will be interpreted as 'twice as good'. Another evaluator giving the same two options '6' and '3', respectively, is saying the same thing, 'twice as good'. Their votes can be combined successfully. If, in the end, a source selection team gives one alternative twice as many 'points' as another, the team has found that one alternative is twice as good as the other.

The Analytic Hierarchy Process (AHP) generates ratio scale data as a natural outgrowth of its paired comparison judgment process. In cases where an evaluation team must rate options against a standard or ideal, the values on the scale being used (whether they are numbers or rating adjectives) have weights that are ratio scale data. With ratio scale data, the results tell us if alternatives are close or far apart. We also know where the candidates or options differentiated one another and by how much. We can explain - in words and in numbers - why a particular bidder won.

Using ratio scales allows us to design a hierarchy of evaluation factors and subfactors tailored to a given selection scenario, to assess performance against these subfactors, and to combine subfactor performance into an overall score that is meaningful. If we use ratio scales, we don't have to worry about assessment-based protests. Simply put, "a decision method that produces ratio scale numbers is the most flexible and accurate."<sup>5</sup>

## An Example

In a U.S. Navy procurement setting described more fully in *Acquisition Review Quarterly*<sup>6</sup>, the objective was to select the best from among 7 alternative ship designs. There were 17 design and performance criteria which were weighted on a scale from 1-17 based on criticality, with the most critical factor given a '1' and the least important factor given a '17'. Further, each alternative was rated from 1-7 regarding how much they possessed a given characteristic or how well they performed against a criterion, with '1' being the best and '7' being the worst.

CRITICALITY FACTORS			DESIGN OPTIONS						
			A	B	C	D	E	F	G
EVALUATION CHARACTERISTICS	AC	15	1	2	6	4	5	3	7
	AL	5	1	2	6	4	5	3	7
	EN	11	1	7	4	3	5	2	6
	CR	6	5	7	3	2	4	6	1
	MA	3	3	4	5	1	6	2	7
	DO	1	1	3	5	4	6	2	7
	ON	8	6	5	3	2	7	4	1
	CN	7	1	3	5	4	6	2	7
	MR	14	7	6	2	4	3	5	1
	EF	2	1	2	6	4	5	3	7
	AD	17	1	2	5	4	6	3	7
	SA	12	1	2	5	4	6	3	7
	RE	9	1	7	4	3	5	2	6
	OF	10	7	6	3	4	2	5	1
	LU	13	1	2	6	5	7	4	3
	DC	4	7	2	5	4	3	6	1
	DS	16	6	7	3	4	2	5	1
TOTAL			471	642	658	568	727	559	651

Scores for each alternative with respect to each criterion (from 1-7) were then multiplied by the criteria weights (from 1-17), and the totals for each ship design were used to prioritize the design options.

In this analysis, the ship design with the least ‘points’ was determined to be the winner. This procurement award was overturned.

Here are the errors committed in this source selection example:

- The 1-17 scale for criteria weights is actually an ordinal scale. The criteria were placed in rank order using this system, and the ranks contain no meaning regarding relative value or importance. The 2<sup>nd</sup> place criterion (designated with a ‘2’) is not necessarily half as important as the 1<sup>st</sup> place criterion (designated a ‘1’).
- The 1-7 scale for alternative performance measurement is, at best, an interval scale, and more likely is an ordinal scale. A score of ‘2’ on this scale is not twice as good as a ‘4’; it is merely two places higher on a 7 point scale.
- Criteria weights (ordinal data) were multiplied by performance scores (most likely ordinal data), resulting in meaningless data. Even if the performance scores and criteria weights were regarded as interval data, the resulting product of interval numbers would be meaningless.

How often is this type of erroneous assessment being utilized in practice? More often than you might think.

## EXAMPLES OF ASSESSMENT BREAKDOWNS

Many organizations have developed standard systems for evaluating bidder performance in source selection. Although we cannot present an exhaustive analysis of them here, we can describe some standard approaches. Our intent is to illustrate a variety of assessment techniques and the issues encountered in using them.

### U.S. Army Materiel Command

“Rating systems which use adjectives or colors are usually the most successful because they allow maximum flexibility in making the tradeoffs among the evaluation factors. A narrative definition must accompany each rating in the system so that evaluators have a common understanding of how to apply the rating. For example, a rating of excellent (or blue or 90-100) could be defined as meaning an outstanding approach to specified performance with a high probability of satisfying the requirement. What is key in using a rating system in proposal evaluations, is not the method or combination of methods used, but rather the consistency with which the selected method is applied to all competing proposals and the adequacy of the narrative used to support the rating.”<sup>7</sup>

#### ASSESSMENT BREAKDOWN:

Unfortunately, the particular “method or combination of methods” is central to effective source selection. As has been described earlier in this paper, some methods – while systematic – are simply erroneous. It is not enough to define performance levels carefully. They must be measured in a valid way. Performance must be assessed in a manner that produces reliable differentiation between the bidders, and assessment must be made against factors whose relative weights are proportional, i.e., numbers on a ratio scale.

### NASA

“(a) Typically, NASA establishes three evaluation factors: Mission Suitability, Cost/Price, and Past Performance. Evaluation factors may be further defined by subfactors...

(b) Mission Suitability factor.

(1) This factor indicates the merit or excellence of the work to be performed or product to be delivered. It includes, as appropriate, both technical and management subfactors. Mission Suitability shall be numerically weighted and scored on a 1000-point scale.

(2) The Mission Suitability factor may identify evaluation subfactors to further define the content of the factor. Each Mission Suitability subfactor shall be weighted and scored. The adjectival rating percentages in 1815.305(a)(3)(A) shall be applied to the subfactor weight to determine the point score.”<sup>8</sup>

#### ASSESSMENT BREAKDOWN:

The adjectival rating percentages referenced above in NASA FAR Supplement 1815.305(a)(3)(A) are as follows:

ADJECTIVAL RATING	PERCENTILE RANGE
Excellent	91-100
Very Good	71-90
Good	51-70
Fair	31-50
Poor	0-30

The Percentile Range above appears to be an interval scale, which introduces practical difficulties (which are referenced in the preceding section of this paper). However, the subdivision of factors into subfactors for Mission Suitability introduces the requirement that factor weights be ratio scale numbers. If factor weights are merely interval scale numbers, then there is a sound basis for protest of awards made using this approach.

### U.S. Air Force

“The color rating depicts how well each offeror meets the evaluation standards. Color ratings are not summarized above the factor level, i.e., factor color ratings shall not be rolled up to the area level...

COLOR	RATING	DEFINITION
Blue	Exceptional	Exceeds specified performance or capability in a beneficial way to the Air Force and has no significant weakness
Green	Acceptable	Meets evaluation standards and any weaknesses are readily corrected
Yellow	Marginal	Fails to meet evaluation standards; however, any significant deficiencies are correctable
Red	Unacceptable	Fails to meet a minimum requirement of the RFP and the deficiency is uncorrectable without a major revision of the proposal

Use of numerical weights is discouraged because it implies that the technical team can differentiate between small differences in technical merit. Such determinations may be extremely difficult to support. Therefore, numerical weighting of evaluation criterion is not recommended.”<sup>9</sup>

### ASSESSMENT BREAKDOWN:

Our experience has shown that it is, in fact, possible for a team to differentiate between small differences on any factor being judged if the right measurement method is used. It does not matter whether the factor is tangible or intangible, whether the data is soft or hard.

Without numbers it is, of course, impossible to perform a ‘roll up’ or summarization of a bidder’s performance. How can ‘best value’ be determined in such a system? It is very difficult to make an overall recommendation when the results look like this:

CRITERION	BIDDER A	BIDDER B
1	Blue	Green
2	Green	Yellow
3	Yellow	Yellow
4	Yellow	Green

Blue=Outstanding, Green=Adequate, Yellow=Marginal, Red=Unacceptable

How much better is a 'blue' than a 'green'? Is their relative importance different depending on the criterion being used? When you roll it up, who wins? And how can it be explained to the losing bidder?

We'll need numbers for that, so let's assume that colors or adjectives are values on a scale of some kind and assign values from 1-4 for the colors above. Assuming the criteria to be equally weighted, this would give Bidder A, the 'winner', 11 points while Bidder B would get 10 points. However, a 1-4 scale is an ordinal scale whose values are based solely on rank, and as we have discussed, arithmetic operations on ordinal data are not valid. Using an ordinal interpretation of these colors or grades would be a sound basis for protest. What if, instead, we assume that the colors are rated on an interval scale like NASA's adjective ratings? Then we must ensure that the criteria are given proportional, ratio scale weights to avoid creating a sound protest opportunity.

Some people argue that it is better to keep the process completely subjective without any weights, but this approach can waste evaluators' time because their analysis is overlooked in the final synthesis of results. Perhaps the Air Force discourages the use of numerical weights because it has been 'burned' by using inappropriate number scales in the past. However, discouraging the use of numerical weights "throws the baby out with the bath water" since numbers are required to explain how much better one value is than another. An audit trail - both narrative and numeric - is necessary to justify any selection.

## **NAVSEA**

"For most source selections, a rating scheme is used to translate narrative evaluations (which highlight strengths, weaknesses, deficiencies, and risks with respect to each award factor) into some sort of numerical scoring...

Once a proposal rating breakdown has been established by assigning rating adjectives to subfactors (and to factors), the task becomes a straightforward matter of assigning a numerical scoring range to each adjective. In this way, a numerical score may be assigned to each subfactor and factor of the proposal.

Typical numerical rating ranges are 0-10 in increments of one; 0-100 in increments of 5, 10, or 20; and 0-1,000 in increments of 50 or 100. Other ranges or combinations may be used to suit the particular acquisition program. Complex weapon system acquisition programs may require multiple rating ranges for different factors/subfactors. The key is the development of a simple method that provides the required discrimination between competing proposals."<sup>10</sup>

## **ASSESSMENT BREAKDOWN:**

We have found that many problems in source selection assessment are based on improper factor and subfactor weights. At the heart of this issue is the difference between 'assigning' and 'deriving' scores and weights. Factor and subfactor weights should be derived by careful comparison to one another. "Which is more important, cost or past performance, and by how much?" is a question that will result in valid, ratio scale information. Assigning weights is as unreliable as assigning scores on a 1-10 rating scale. It is difficult, if not impossible, to force all evaluators to use the same scale the same way.

As we have discussed earlier in this paper, using a scale consistently won't help if it's the wrong kind of scale. A ratio scale must be used so that assessments on different factors and, perhaps, made by different evaluators can be combined meaningfully. In this case, the scales being suggested are interval scales. What if the criteria weights are assigned in this same manner? In that case, we would be multiplying interval level assessments for each bidder's performance times interval level criteria weights, resulting in meaningless data. A source selection conducted in this way would be subject to protest.

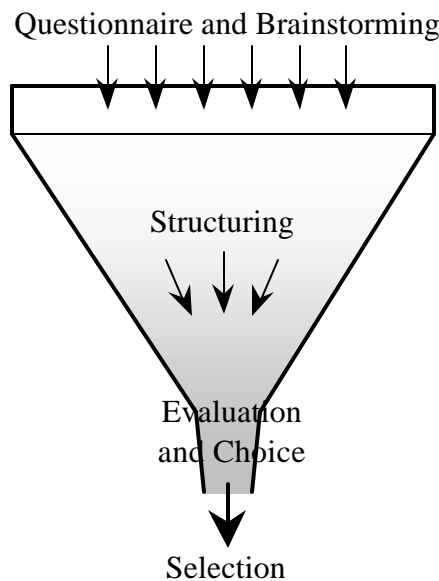
## LINKING ASSESSMENT TO EVALUATION

Our methodology for source selection is based on the use of Team Expert Choice™ (TeamEC™) and its Analytical Hierarchy Process (AHP) foundation. This approach is logical, rational, and mathematically valid, yet it is easy to learn and use. It is also flexible and works equally well with tangible and intangible requirements. Before we describe in detail how this process works, it is important to keep several things in mind:

1) TeamEC™ is a general purpose tool which can support many different kinds of evaluation processes outside of source selection. In this section, however, we will focus on providing you with sample screens and descriptions that are exclusively about source selection.

2) TeamEC™ is NOT an electronic source selection (ESS) evaluation software product. The purpose of ESS tools is to capture comments of individual evaluators and keep them organized in an electronic format. They are very powerful databases that can be useful in keeping a qualitative record of a proposal evaluation. They are often used in conjunction with TeamEC™, but they focus on Recording rather than on Measurement processes.

3) As we have said previously, we can ‘plug into’ the source selection process and add value at almost any point because of the flexibility of TeamEC™. However, we will assume in this narrative that our methodology is used from the outset and throughout the process.



### Criteria Development

Our first step is to get the key individuals (source selection officials, evaluators, or both) together for a criteria development session. While it is possible to collect some preliminary criteria information asynchronously, it is essential that criteria be finalized in a same time setting. (Face-to-face or same time/same place is strongly preferred to same time/different place mode for this step.) Using

TeamEC's™ Questionnaire & Brainstorming module to develop an initial list of factors, the team might create a list such as this:

The screenshot shows a software window titled "Questionnaire & Brainstorming C:\TEAMEC\SESSIONS\WHITEP~1\SELECT.BST". The window contains a table with the following data:

Identify most important source selection criteria	
All Items sorted by Overall Score, DESC	Rate 1-10
Technical Aspects	9.40
Cost	8.90
Personnel	8.89
Commitment	8.70
Past Performance	7.90
Flexibility	7.20
Statement of Work	6.90
State of the Art	6.20
Management	6.00
Insight	6.00

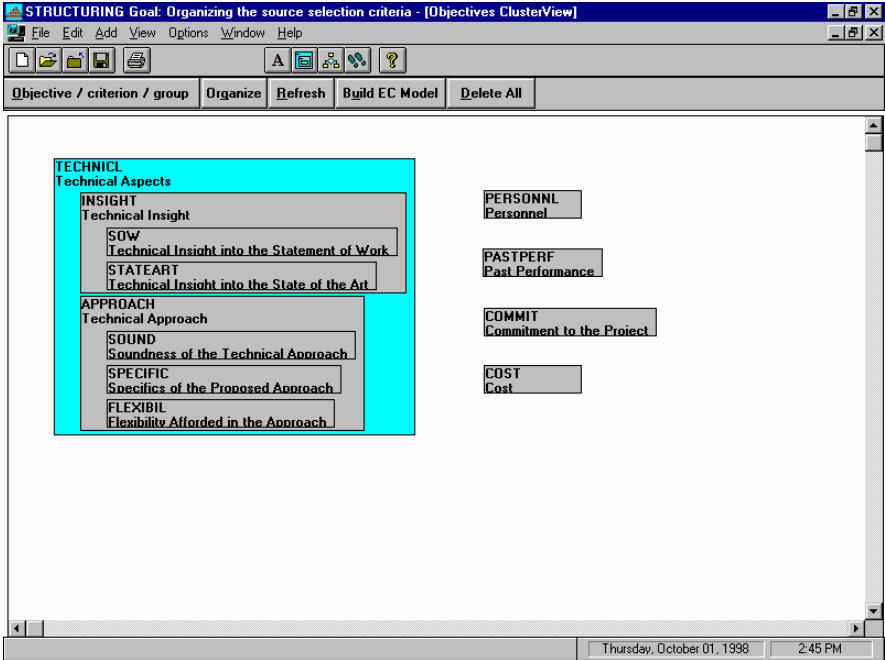
The status bar at the bottom of the window displays: Ready | Brainstorming | Keypads <On> | COM1 | Wave 1 | Pg 1 | 01-Oct-98 | 3:34 PM

Criteria development, however, is more than simply brainstorming a list of potential factors to consider. It is the process of identifying and structuring criteria that will be used to evaluate proposals, and it differs from requirements that are in the RFP in several ways. First, the term 'requirement' traditionally refers to specific criteria that, if not met in the proposal, will exclude an offeror from winning a bid. These are often among the first things to be written into an RFP. Evaluation criteria, on the other hand, are general criteria that assess the value added to the purchasing organization. They are often a mix of intangibles like innovation and corporate culture, with tangibles like system performance and cost.

One of the most unnecessarily time-consuming parts of the source selection process is getting a final list of criteria that all the evaluators understand and can apply consistently to evaluate offerors. For this critical step, we use TeamEC's™ criteria development utility, Structuring. A list (such as the one above) can be automatically transferred into the Structuring module, or the group can generate potential factors on-the-spot guided by a facilitator who captures the team's ideas by entering them directly. Once the team has a preliminary list of criteria, the group begins an organizing process we call clustering.

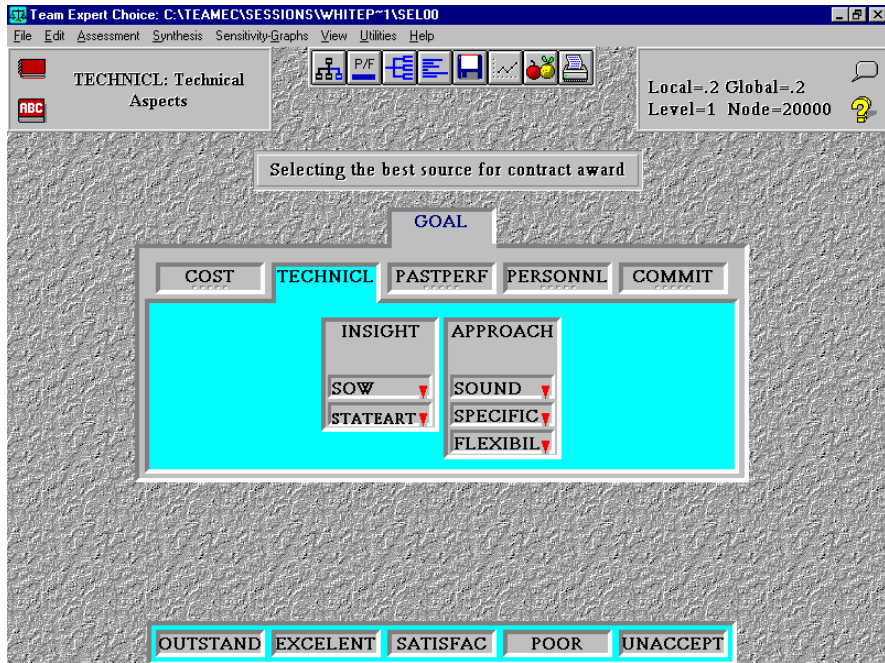
Clustering involves asking members to group ideas into "theme categories" which eventually become the evaluation hierarchy. They look at a common computer display that is projected onto a screen, and they suggest, verbally, which ideas fit together. At times, a team's discussion of a set of potential criteria will surface new evaluation factors, which are added into the mix. The facilitator, usually assisted by a technographer, drags & drops the common concepts onto one another, linking

them together. Those who are familiar with Total Quality tools and techniques will recognize this Structuring process as Affinity Diagramming:



Categories can drill down several levels so very complex models can be built. Categories often act as a guide to helping evaluators identify issues that they might have missed in the brainstorming phase because they help evaluators organize related criteria.

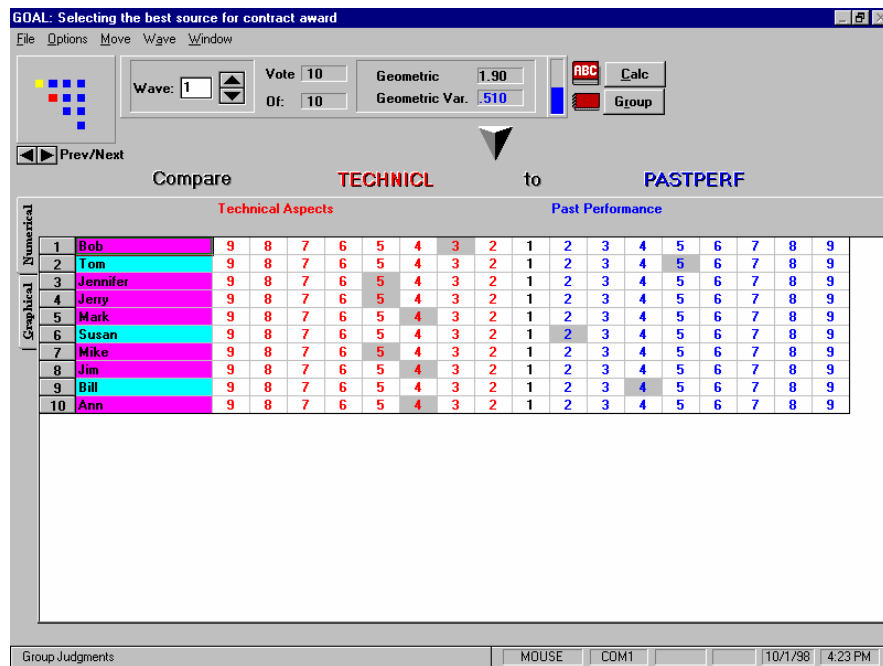
Once the evaluators have finished their criteria development and organization process, TeamEC™ automatically expands the clusters into a hierarchy ready for prioritization:



## Deriving Criteria Weights

Deriving criteria weights is accomplished by comparing the evaluation criteria to one another in a pairwise fashion, measuring proportionality by using a Ratio Ruler. The middle point on this ruler indicates that the two factors being compared are equally important or preferable. This pairwise approach allows evaluators to compare tangibles to intangibles on a reliable scale.

Each evaluator expresses an opinion using a wireless radio frequency keypad, and all individual judgments are collected and aggregated automatically into a group judgment. A sample comparison of two evaluation factors is shown below:

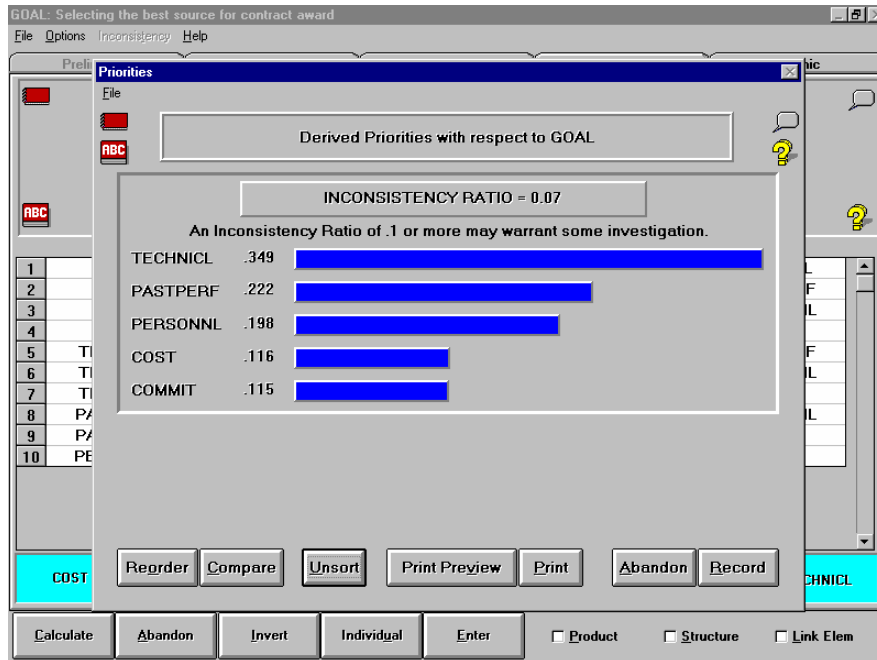


## Consensus at the Criteria Level

The use of wireless keypads for assessment significantly reduces the amount of time required to gather feedback from the team. The larger the group, the more powerful this benefit will be. Further, a team will see where they already agree. Time can be saved for areas of disagreement, which are highlighted graphically.

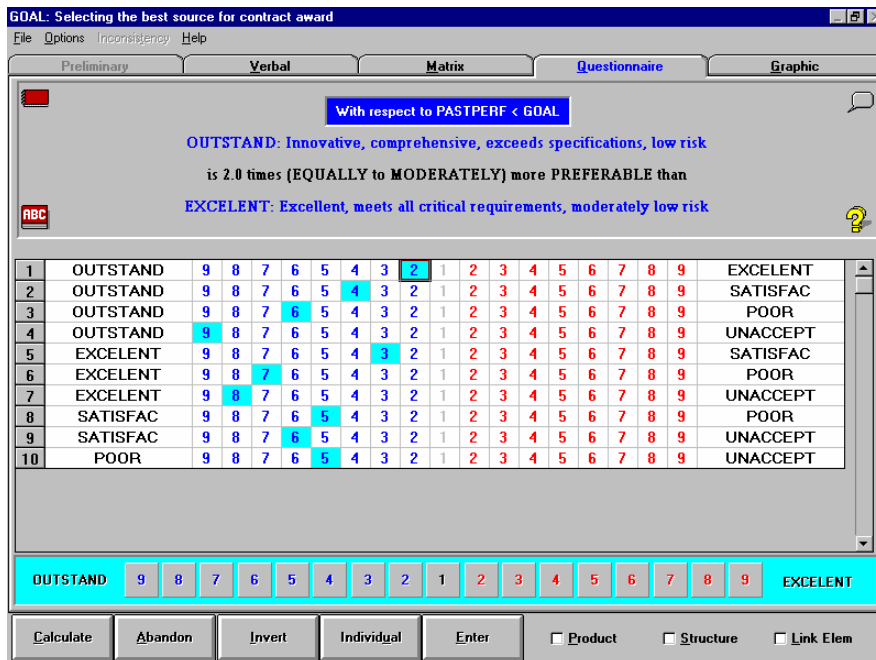
This real-time criteria comparison process is very powerful because people are asked to justify why they feel Criterion A is moderately more important than Criterion B, for example. This helps all evaluators understand the problem better because they get new information from one another and clarify stakeholder positions. For instance, the engineer will better understand what the financial analyst thinks the criteria mean and where his priority falls in the decision. This helps the engineer to be more effective at communicating with the financial analyst because both people have already clarified their positions.

After all evaluation criteria have been measured against one another, TeamEC™ derives the overall criteria weights. These proportional weights are displayed to the team both graphically and numerically:



## Developing Rating Scales

In TeamEC™, rating scales are developed by the evaluators so that they can give offerors the correct numerical score for a rating that is used to evaluate each vendor with respect to an objective. Let's take a qualitative scale: Outstanding, Excellent, Satisfactory, Poor, and Unacceptable. Evaluators are asked to compare Outstanding to Excellent, Excellent to Satisfactory, Satisfactory to Poor, and Poor to Unacceptable in a pairwise fashion:



The results are numbers on a ratio scale, whose values might look like this:

Outstanding = 100  
 Excellent = 72  
 Satisfactory = 37  
 Poor = 15  
 Unacceptable = 8

If these were the results, the evaluators would be saying that they give a lot of value to an Outstanding approach, and substantially less to an Excellent. Now when evaluators give ratings to offerors, the numerical scores will mean quantitatively what the evaluators mean qualitatively when they submit their scores.

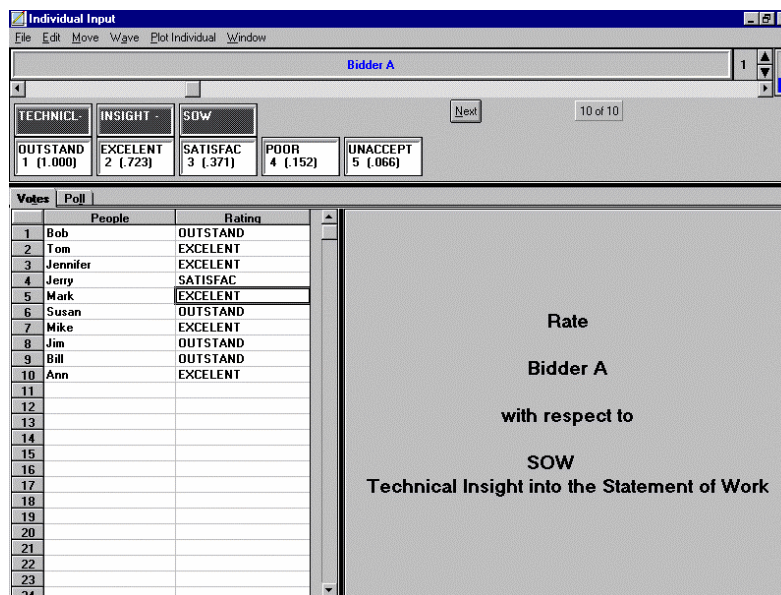
In many cases, the same scale has a different meaning depending on what is being measured. For example, 'Outstanding' Past Performance might be valued as twice as good as 'Excellent' Past Performance, while an 'Outstanding' Statement of Work might be seen as only marginally better than an 'Excellent' Statement of Work:

RATING	PAST PERFORMANCE	STATEMENT OF WORK
Outstanding	100	100
Excellent	50	90

Throughout the evaluation process, the ratings scales and their associated weights or values can be customized to fit what is being measured.

### Evaluating Bidder Performance

Once criteria and rating scale intensities have been derived, bidders can be evaluated using the scale (or scales). A facilitator then leads the group step-by-step through the evaluation process. At each point, the group uses the agreed-upon scale to measure bidder performance for each factor or subfactor:



## Consensus at the Bidder Evaluation Level

As with criteria comparisons, the evaluation team sees where it agrees and disagrees on a bidder evaluation. Where there is no disagreement, we move on. Where there is disagreement, we can spend time discussing reasons for different judgments, which might include misinterpretation of performance level definitions or disagreement on bidder performance, for example. This discussion is the most valuable aspect of the use of our methodology since it brings a team to a common, unified view of the entire evaluation. Because we save so much cycle time in other areas of the assessment process, we have the time to devote to discussion, clarification, and consensus where needed.

Some teams adopt an approach that calls for preliminary votes to be collected using keypads, which record individual judgments or perspectives as a precursor to discussion. Then, after discussion, they agree to enter a consensus judgment for the team – rather than individual judgments – as their final assessment. With our approach, the facilitator can do this easily on behalf of the group.

When evaluators come to consensus using our method, they are equipped with the tools (clear definitions, factor weights, rating scales, and bidder assessments) to roll-up individual evaluations effectively. The evaluators all understand the criteria because of our advanced criteria development process. They also understand the quantitative differences between ratings so that they give meaningful numerical scores to offerors. This helps to assure that they treat all offerors fairly when they are evaluated - a major factor in protecting against protests.

## Synthesis

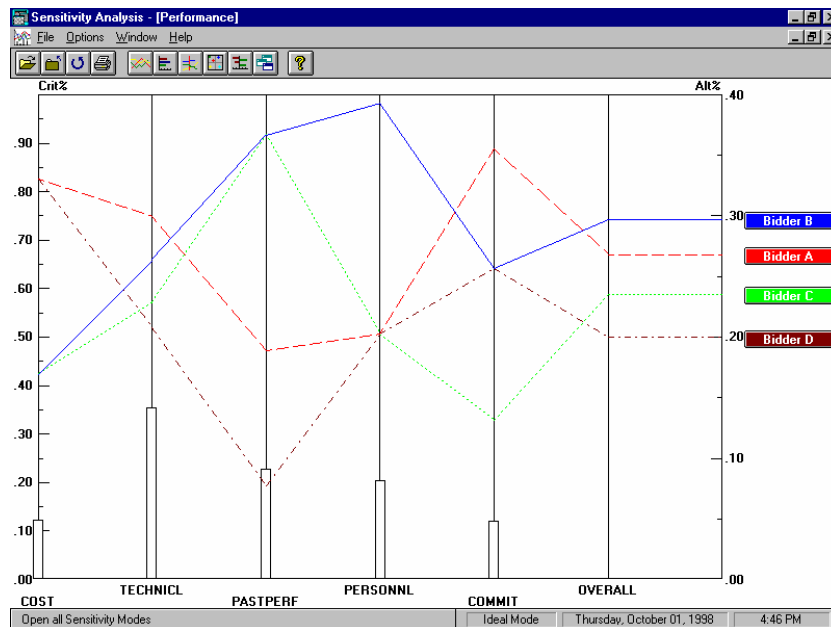
The process of synthesis or the ‘rolling up’ of results is carried out automatically by TeamEC™. Because we break a large, complex problem into small, manageable pieces and evaluate each piece on a meaningful scale, we are able to develop overall results that accurately represent the final scores of the vendors. If one vendor scores .67 overall and another scores .62, the winning vendor should be 5% better:

Alternatives	TOTAL	APPROACH-SOUND	SPECIFIC	FLEXIBIL	PASTPERF	PERSONNL	COMMIT
		0501	0501	0501	.2220	.1901	.1148
1 Bidder B	0.665	EXCELENT	SATISFAC	SATISFAC	EXCELENT	EXCELENT	EXCELENT
2 Bidder A	0.617	EXCELENT	SATISFAC	EXCELENT	SATISFAC	SATISFAC	OUTSTAND
3 Bidder C	0.521	SATISFAC	EXCELENT	EXCELENT	EXCELENT	SATISFAC	SATISFAC
4 Bidder D	0.460	EXCELENT	OUTSTAND	POOR	POOR	SATISFAC	EXCELENT
5							
6							
7							
8							
9							
10							

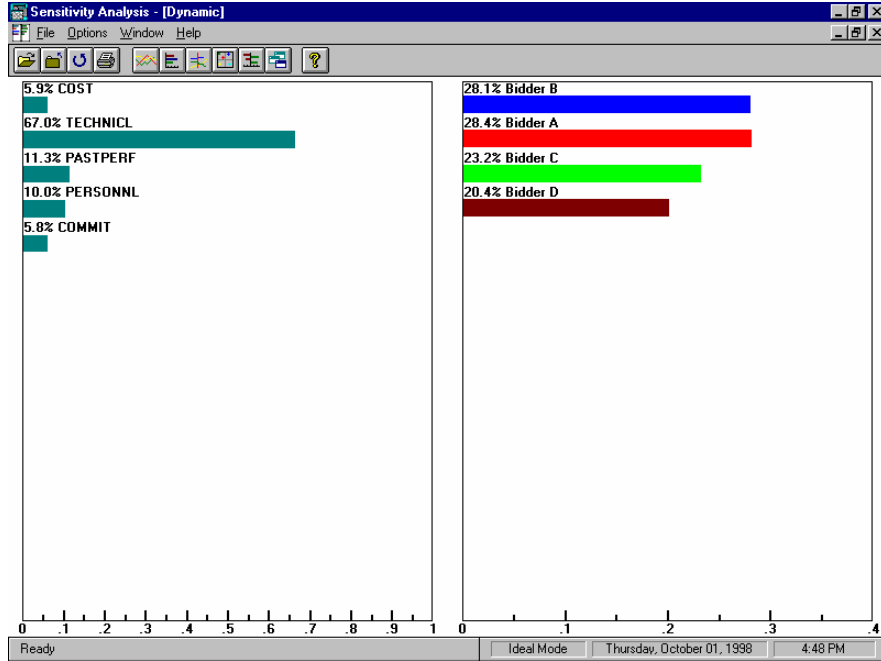
If the group has used an ESS tool to capture justifications for the ratings, the result is a rock solid numerical and qualitative evaluation of vendors.

## Sensitivity Analysis

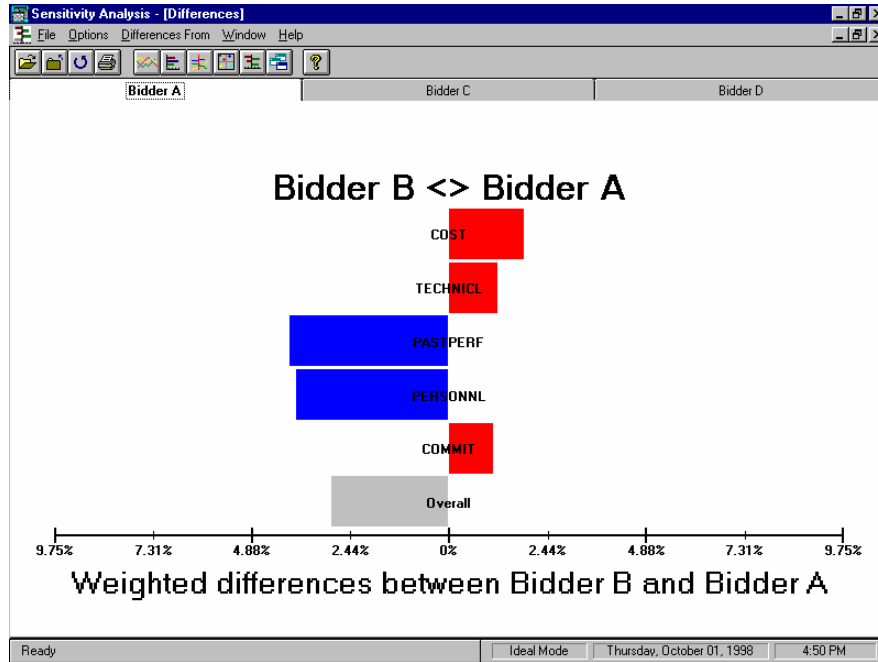
Having a solid assessment approach is very important, but the ability to explain it to a losing bidder is critical. Sensitivity Analysis permits the evaluation team to anticipate questions and challenges, to see how sensitive their selection is to changing scores at various levels. Performance Sensitivity (below) gives a top level view of how each factor was weighted and how each bidder performed with respect to them:



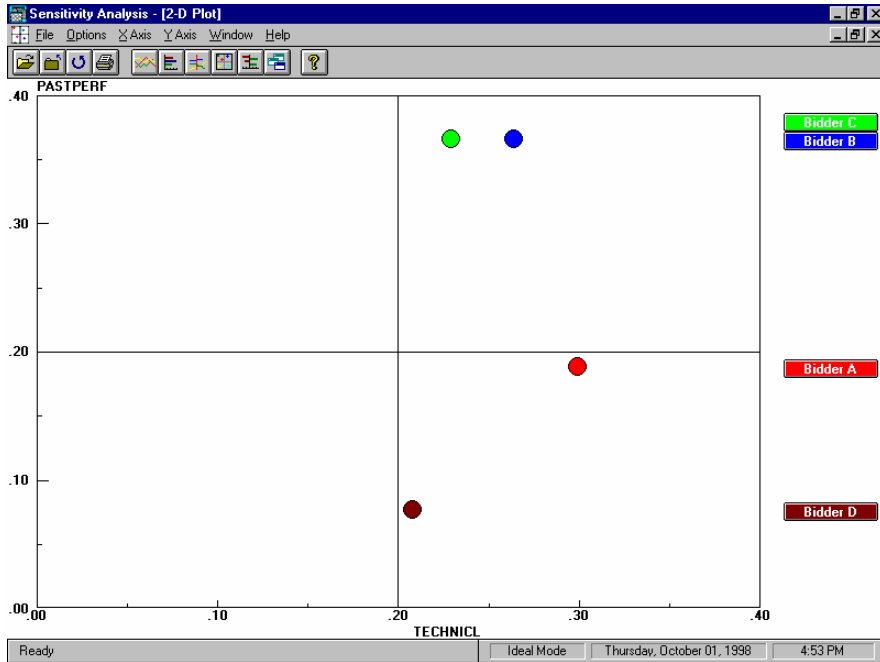
How valuable is it to be able to say to a bidder that his challenge doesn't affect the overall outcome, even if it were accepted as the evaluation team's rating? What Ifs can be simulated on screen to examine if the selection is sensitive to changes in a given area using Dynamic Sensitivity. In the example shown below we illustrate (by dragging the 'Technical Aspects' bar to the right) that 'Technical Aspects' would have to be twice as important for Bidder A to catch up to Bidder B and win the award:



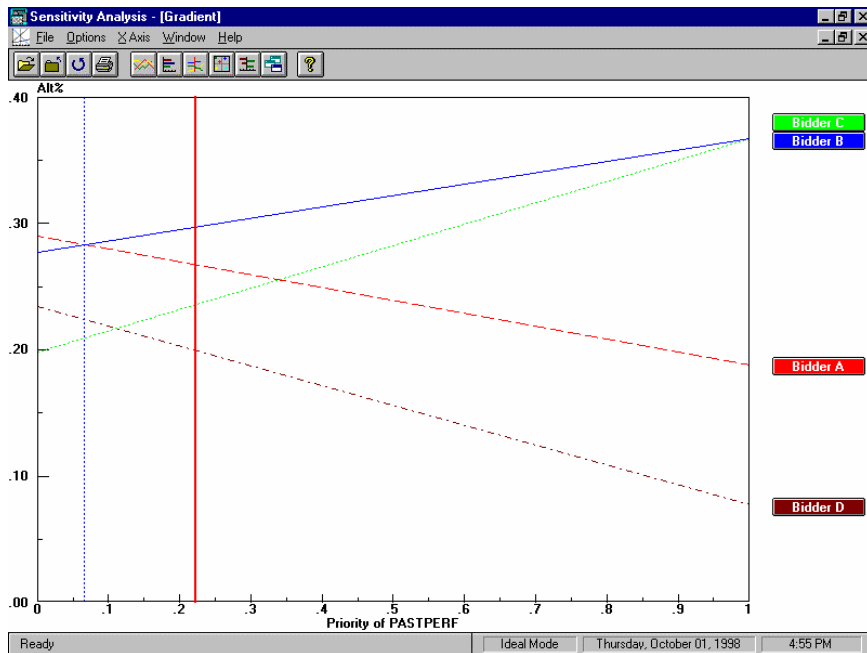
Difference Sensitivity graphically illustrates the head-to-head performance of a losing bidder versus the winner:



2D Charts are used to graph any two factors - 'Cost' and 'Technical Factors' for example – and all bidders' scores in quadrants. In the example below, you can easily see that Bidders A and B are rated closely on 'Technical Aspects', but the evaluation team found Bidder B to be almost twice as preferable in terms of 'Past Performance':



Gradient Analysis supports factor level investigations. In this example, we illustrate that, in order for Bidder A to overcome its low 'Past Performance' rating, that factor would have to be reduced in priority by over 100%:



Sensitivity analysis of the preliminary selection or recommendation will pinpoint where more careful data collection and further investigation is needed. It will focus evaluation team efforts in areas

where time should be spent, i.e., on issues to which the selection is highly sensitive. Conversely, it will also identify factors to which the selection is relatively insensitive. It is particularly effective at helping to explain to a losing vendor that a higher score on one factor or another might not have any effect on the overall selection.

## SOURCE SELECTION CASE STUDY

The following is a description of the Information System Product Evaluation Methodology, a copyrighted work of Birch & Davis Associates, Inc.

---

When evaluating clinical information system products, B&D's method of choice is the DoD's Commercial Off-the-Shelf System Evaluation Technique (COSSET). This technique represents all stakeholders equally, minimizes bias and conflict of interest, and provides a formal means to address end-user's concerns. It provides a systematic, fair, and objective approach to narrowing the universe of candidate products to those few that realistically may meet or exceed the customer's product requirements. Most organizations know up front which products will meet their threshold needs. The COSSET simplifies the dilemma of determining which product, or products, will provide the best satisfaction to end-users and stakeholders at the least risk to the organization. The COSSET process answers the following question: "*How well* does each product satisfy the organization's requirements?"

The COSSET process includes six main steps as follows:

1. Work group formation
2. Functional analysis/requirements development
3. Product evaluation tool-set development
4. Preliminary vendor survey
5. Candidate product identification
6. Product evaluations

The COSSET uses a number of processes, and techniques to achieve its results. Work group teams, ranging from 5 to 15 members, use the Analytic Hierarchy Process (AHP) through Team Expert Choice™ software to develop three of the product evaluation tools and to score the product evaluations. Team Expert Choice™ is used because it is a flexible, easy to learn, decision-making tool for complex, multicriteria problems, particularly where both the qualitative and quantitative aspects of a product need to be considered. Team Expert Choice™ has proven to be a valuable resource as it allows users to trace global decisions directly back to an individual requirement, regardless of the number of requirements. In one evaluation, for example, more than 100 requirements were involved. As a result, this process is ideal for building or migrating to enterprise systems.

After the conclusion of the product evaluations, a level of performance is computed for each product. Results of the product evaluations are based solely upon quantitative and qualitative inputs from the evaluators. The report identifies each product's strengths and weaknesses, by comparing each product against a model of the desired system's functional preferences and by analyzing the product's ability to satisfy the organization's requirements. The analysis can be high-level; based on main features and capabilities, or can extend to any level of granularity necessary to illuminate which product provides the best value. Based on the findings, the work group can make informed recommendations and the decision-makers team can determine rational next steps (i.e., continue refining requirements, build a system, progress to acquisition).

To date, the COSSET has been used over a period of 18 months and 32 clinical information systems products have been evaluated including Ambulatory Care, Emergency Department, Immunization Tracking, Dental Imaging, Resource Management, and Mammography Tracking. Based upon the findings of these evaluations, decision-makers are making informed, apolitical acquisition decisions about available products. Depending on how fast the organization wants to proceed, the process can take just a few weeks or months.

Lessons learned indicate that this methodology:

- Integrates functional analysis, requirement definition processes, and product evaluation methodology into a business process that can be completed as one or several initiatives.
- Integrates organizational and technological considerations with subject matter expert findings to enable informed decisions.
- Sets clear guidelines and criteria for managers to follow when identifying and defining requirements, setting priorities, and evaluating information systems.
- Evaluates large systems having hundreds of requirements just as easily as small systems having fewer requirements. The process can evaluate vertical market software just as easily as horizontal-market software.
- Facilitates quality in Request for Proposal (RFP) development, contract award negotiations, and debriefing sessions to participants.
- Reduces political tensions surrounding complex decision-making and enables end-users to communicate effectively with decision-makers and colleagues.
- May produce a library of requirements, scenarios, surveys, and product evaluations that can be reused or combined with future evaluations.

When using the COSSET, B&D's product evaluations enable end-users to be effectively involved throughout the system development process. The process successfully manages the complexity when migrating to enterprise systems and produces rapid, tangible, and proven results. Stakeholders from all levels of the organization can reach a consensus on identifying requirements, setting priorities, and recommending effective information system solutions.

## CONCLUSIONS

As we have shown, a variety of assessment methods are used by industry and government to support source selection processes, but many have strong potential as the basis for a protest. The numbers or scores generated by these different methods are not equally valid, leading some procurement groups to discourage the use of numbers altogether, most notably the U.S. Air Force. However, assessment scores are essential for determining the relative priorities of selection criteria and for measuring performance against them. Further, assessment scores are vital in the justification process since they measure the degree of separation between suppliers or vendors on each criterion individually as well as overall.

Assessment scores are also critical as feedback to suppliers or vendors since they indicate how well they performed versus their competitors or areas where they improved compared to prior submissions. Learning is facilitated in an environment which uses assessment scores within the framework of a structured, disciplined process. Evaluator prejudice and bias can be removed, enhancing a team's ability to communicate the outcome and gain acceptance by bidders.

Our source selection methodology is uniquely effective in that:

- Our assessment methodology is immune from protest.
- We can reduce criteria development time from 4-6 weeks to a matter of days or hours.
- We make the integration of subjective judgments and objective data a reality.
- Through our simultaneous judgment process, areas of obvious agreement on assessments are identified early, and time is not lost on them. As a result, evaluators are able to invest time on areas of disagreement, a discussion process that leads to clarity and consensus.
- Different stakeholders, evaluators, or subject matter experts can be involved in the decision, providing their judgments in different parts of the hierarchy.
- We help you develop a rationale, a justification that you can use – in real time - to explain and defend your selection.
- We can adapt to any standardized assessment procedure (adjectival, color, performance level, etc.) and introduce number scale validity to the process.

Our technique – based on TeamEC™ and its Analytic Hierarchy Process (AHP) - generates valid assessment scores every time it is applied. It is also unique in its ability to measure intangibles alongside tangible factors in a selection scenario. In fact, it is ideally suited to accommodate current issues and new directions in source selection. For example:

## **Lowest Price Technically Acceptable**

While it is true that no trades between cost and non-cost factors are permitted under these conditions, our approach solicits the best subjective judgments, synthesizes these with objective factors, and permits an assessment against a technical baseline of minimum acceptability. It should be noted that - subject to some limits - Past Performance, which frequently involves subjective factors, can be used in this approach.

## **Best Value**

Best value is the most advantageous offer, price and other factors considered, providing the best mix of utility, technical quality, business aspects, risks, and price for a given application. Our approach provides the best process for assessment of non-cost factors and produces a meaningful relative numeric ranking which can be combined with cost data to yield a Benefit/Cost ratio ranking to determine a Best Value offeror for the government.

One significant non-cost factor is Past Performance, a factor that can be used to represent the risks - programmatic, technical, cost, validity, and quality - associated with a contractor. By its nature, there will be subjective assessments within Past Performance ratings. A number of case studies on the use of Past Performance ratings in Best Value source selections have indicated how valuable it can be to the process.<sup>11</sup>

Frequently, Past Performance is weighted between 25 and 50 percent of all non-cost factors. It is the most likely and significant candidate for the integration of subjective and objective subfactors and can, therefore, benefit greatly from the use of our approach. Proper implementation of Past Performance assessment requires the relative ranking of offerors. Too often, however, relative ranking is performed improperly, and we predict that without the use of our approach, the continuing emphasis on past performance will increase the potential for assessment method errors as a basis for protests.

## **A Call to Action**

We can assist you in linking proper assessment methods to your source selection efforts, making your selections easy to explain and defend. When you are ready to implement protest proof source selection, call us at 1-800-447-0506 or send an e-mail to [selection@expertchoice.com](mailto:selection@expertchoice.com).

## BIBLIOGRAPHY

1. "OSD - Notifying and Debriefing Unsuccessful Offerors", Best Value Handbook, David Drabkin, ODUSD (AP&P).
2. "AF- Good Debriefs", Defense Acquisition Deskbook Version 2.4.110, File Owner: KeeKee Schuh, AFMC/AQ.
3. "Statement of Guiding Principles", Federal Acquisition Regulations System, Part 1.102.
4. "On the Theory of Scales of Measurement", S.S. Stevens, Science, (103, 1946), pp. 677-680.
5. "Decision by Objectives (How to convince others that you're right)", Dr. Ernest H. Forman, Part I Section C, Decision-making Concepts & Methodologies, p.27.
6. "Using Data Types and Scales for Analysis and Decision Making", Richard Pariseau and Ivar Oswalt, Acquisition Review Quarterly, Winter 1994.
7. "Contracting for Best Value- A Best Practices Guide to Source Selection", AMC-P 715-3, I January 1998, United States Army Materiel Command.
8. "NASA evaluation factors", NASA Federal Acquisition Regulations Supplement 1815.304-70.
9. "Use of Rating Techniques", Air Force Federal Acquisition Regulations Supplement, Appendix AA, AA-304 (b) and (c), Formal Source Selection for Major Acquisitions, AFAC 96-1, 13 June 1997.
10. "Use of Adjectives" and "Numerical Rating Ranges", NAVSEA Source Selection Guide 4.4.1 and 4.5.6, 12 February 1993.
11. "A Guide to Best Practices for Past Performance", Office of Federal Procurement Policy (OFPP), Interim Edition, May 1995.

Team Expert Choice™ and TeamEC™ are registered trademarks of Expert Choice, Inc.