3 Guidance for Successful Evaluations

In developing STEP, project leads identified several key challenges in conducting technology evaluations. The following subsections address the four challenges identified by MITRE evaluation teams that are critical to ensuring an evaluation's success:

- Methods used to evaluate and score products,
- Communication during the evaluation process,
- Ensuring evaluation integrity, and
- Creating an evaluation timeline.

These challenges were echoed and addressed in several literature searches on decision making. As stated in an article [6] on methods and best practices in evaluating alternatives:

> "There are many potential mistakes that can lead one awry in a task…Some concern understanding the task. Others concern structuring the decision problem to be addressed. Still others occur in determining the judgments necessary to specify the [scores]… These mistakes frequently cascade… 'When this occurs, the [scores] provide little or no insight, contribute to a poor decision, and result in frustration with the decision process."

3.1 Methods Used to Evaluate and Score Products

In a technology evaluation, teams must evaluate and score products against a set of evaluation criteria in order to determine the best choice to meet their sponsor's needs. Teams must produce a clear assessment of the products and provide a rationale that can be used to make and justify decisions. The process involves

1. establishing a set of evaluation criteria and, as appropriate, dividing the criteria among a set of categories,
2. determining a scheme for scoring products against the evaluation criteria
3. providing a set of numerical weights to determine the relative importance of the criteria and evaluation categories
4. computing the overall score for each product

Teams often use a spreadsheet such as the one in Table 2 to track the evaluation criteria, scores, and weights, and calculate the total weighted scores for each product (see Appendix B for this Evaluation Criteria Template).

Table 2: Spreadsheet for capturing evaluation criteria, weights, and scores

#	Evaluation Criteria	Description of How to Test the Criteria	Weight	<product P1 scores>	<product P2 scores>	<product P3 scores>	<product P4 scores>	<product P5 scores>
1.0	Category 1 Title							
1.1	Criteria A	-description-						
1.2	Criteria B	-description-						
1.3	Criteria C	-description-						
1.4	Criteria D	-description-						

The following subsections provide guidance for accomplishing steps 1- 4 above. This guidance comes from the multi-attribute utility (MAU) analysis, within the mathematical field of decision analysis. Decision analysis is concerned with providing a mathematical framework for decision making, so that decision makers can rigorously and consistently express their preferences, in such a way that their results can be readily and logically explained.

Multi-attribute utility (MAU) analysis [1, 2, 3, 4, 5, 6, 7, 10, and 14] is a well-established decision analysis method that specifically addresses how to select one alternative from a set of alternatives, which is akin to selecting a particular product from a set of products in a given technology area. MAU analysis follows steps 1- 4 above to compute the overall score, or utility, of each alternative under consideration. By following the rules and principles of MAU analysis, evaluation teams can perform straightforward, rigorous, and consistent decision making. Furthermore, teams can back up the integrity of their results through an established scoring method that is recognized as comprehensive and fair.

3.1.1 Establishing Evaluation Criteria

In preparing for the evaluation testing, the first step is to establish the evaluation criteria. This is a key step, because at the end of the evaluation, the results will be a reflection of how well the team created their evaluation criteria. In order to generate these criteria, the team should conduct independent research and request guidance on all aspects and objectives of the problem from the government sponsor and subject matter experts. Through this research, the team will ensure that the sponsor's primary needs/wants are addressed, as well as critical functional (e.g. security) capabilities or nonfunctional (e.g., policy, vendor support) issues.

Evaluation criteria should be specific, Boolean (two-valued) types of questions that are clearly stated and can be clearly tested. The following tips are provided for writing individual criteria statements. First, use the "who shall what" standard form to prevent misunderstanding. In other words,

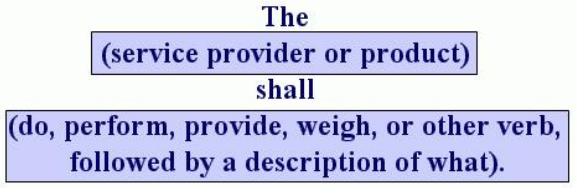

Figure 4: Standard form for writing the evaluation criteria

In writing these statements, avoid the following pitfalls listed in [xx]:

- Ambiguity – write as clearly as possible so as to provide a single meaning
- Multiple criteria – criteria that contain conjunctions (and, or, with, also) can often be split into independent criteria
- Mixing evaluation areas – do not mix design, system, user, and vendor support criteria in the same evaluation category.
- Wishful thinking – "Totally safe", "Runs on all platforms".
- Vague terms – "User friendly", speculative words such as "generally", "usually"

In addition to the evaluation criteria statements, provide a description of how each criterion will be tested. Following these tips will help ensure that each evaluation criterion is carefully written, independent, and clearly states *what* is tested, *how* it is tested, and the desired outcome.

3.1.2 Scoring the Products

The next step is to determine how products will be scored against the evaluation criteria. For example, teams could use the following function u_i:

- $u_i(a_i) = 0$ if a product does not meet evaluation criteria a_i
- $u_i(a_i) = 1$ if a product partially meets evaluation criteria a_i
- $u_i(a_i) = 2$ if a product fully meets evaluation criteria a_i

This function is a constructed scale because each point is explicitly defined. Constructed scales are often useful because they allow both quantitative and qualitative criteria to be measured. Teams may prefer to assign scores based on a standard unit of measure (e.g., time, dollars), a complex function, or another function type.

By convention, in MAU analysis, any scoring function should be normalized so that the scores fall in the range from 0 to 1. Normalizing the above constructed scale gives:

- $u_i(a_i) = 0$ if a product does not meet evaluation criteria a_i
- $u_i(a_i) = .5$ if a product partially meets evaluation criteria a_i
- $u_i(a_i) = 1$ if a product fully meets evaluation criteria a_i

Therefore, in the above example, a product that fully meets a criterion during testing will receive a score of 1, a product that partially meets a criterion will receive a score of .5, and a product that does not meet a criterion will receive a 0 for that item. These are not the only possible scale values. In this case we have a discrete set of three values. We could have a larger discrete set or a continuous set between 0 and 1.

3.1.3 Computing Weights

The final step is to assign weights w_i to each criterion. These weights serve as scaling factors to specify the relative importance of each criterion. Because they are scaling factors that specify relative importance in the overall set of criteria, they should be nonnegative numbers that sum to 1.

There is no "best" method for choosing weights. The choice depends on the principles and axioms that the decision maker wishes to follow, level of detail desired for the weights, and the computing resources available for calculating the weights.

A variety of methods have been proposed for eliciting weights [1, 2, 3, 4, 10, and 14]. These methods include:

- Weighted Ranking
- Analytic Hierarchy Process (AHP)
- Trade-off method (also called Pricing Out)
- Paired comparison (also called Balance Beam method)
- Reference Comparison

These methods are compared in Figure 5 below and the Paired Comparison and Reference Comparison methods are recommended for use by MITRE evaluation teams.

The first three methods, weighted ranking, AHP, and the trade-off method, are not recommended in this guide for the following reasons. Both weighted ranking [2, 9] and AHP [5, 10] are popular methods, but they can be manipulated in ways that result in certain basic logical flaws, and as a result, are often rejected by decision analysts as acceptable methods for computing weights [2, 4, 11, 14]. The Trade-Off method [2, 3, 6] is also a well-accepted method, but is not recommended because of the computational resources required to derive weights for more than 10 alternatives. Several commercial decision software packages are available that implement this method.

The Paired Comparison and Reference Comparison [3, 9, and 14] are recommended in this guide for use by evaluation teams because they are widely accepted and practical to perform by hand. The Paired Comparison is a good choice when deriving weights for 10-100 alternatives. Alternatively, the Reference Comparison method is a good choice when deriving weights for 100+ evaluation criteria. It requires fewer computations than Paired Comparison; however it provides less granular weights.

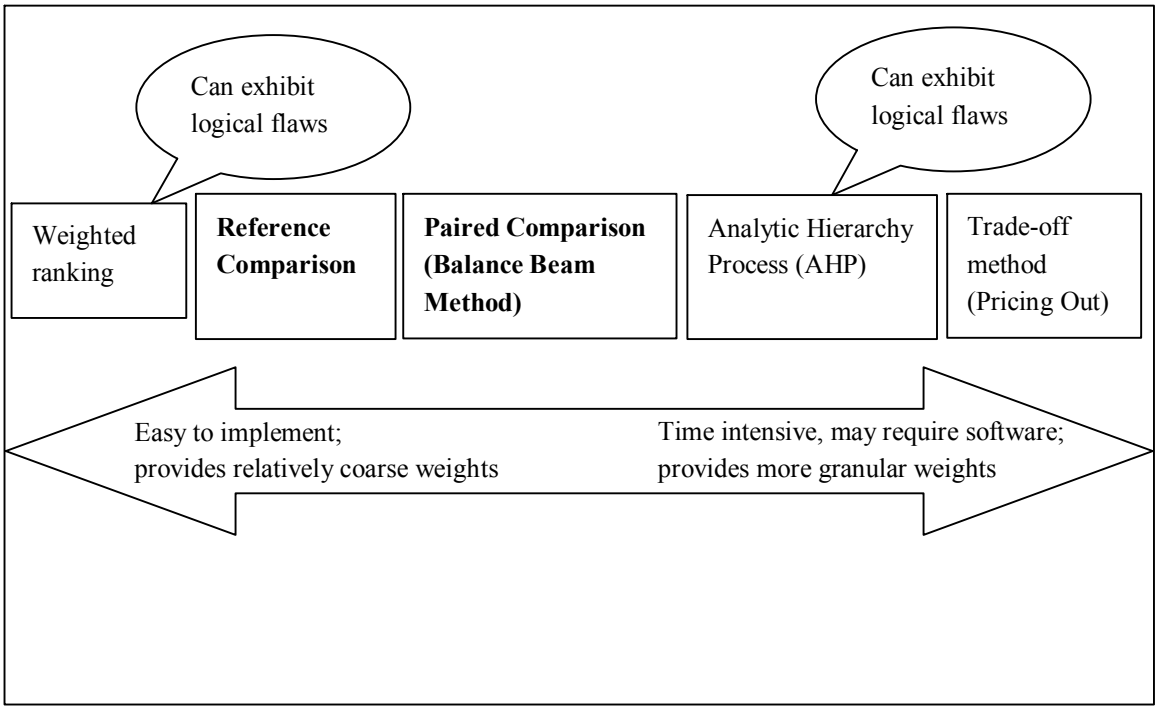

Figure 5: Comparison of Weight Assessment Methods. Reference Comparison and Paired Comparison are recommended in this Guide for evaluation teams

Paired Comparison:

This method is a good choice for deriving weights for 10-100 alternatives and is best explained with an example. Given a set of evaluation categories or a small set of evaluation criteria, determine a basic ordering from highest importance to least importance. Throughout these weight assessment methods, basic orderings and relative importance is decided by the team and will be subjective.

Example:

Most important = A

B

C

D

E

F

Least important = G

For example, in an evaluation of a security product, security is the most important category, followed by auditing, administration/management, and then vendor resources.

Starting with the alternative of highest importance, express its importance with the alternatives of lower importance in terms of a <, =, or > relationship. There is no rule about coming up with this expression, it is determined by the evaluation team. Obtain an equality (=) relationship whenever possible to make it easier to solve the equations at the end. Repeat this with the alternative of next highest importance, until each alternative is expressed in terms of lower-order alternatives, as shown:

Paired Comparisons (Balance Beam Comparisons)	**Relationship**
A < B + C **A = B + D**	**A = B + D**
B > C + D B < C + D + E B < C + D + F **B = C + D + G**	**B = C + D + G**
C < D + E **C < D + F** **C > D + G**	**C > D + G** **C < D + F**
D = E	**D = E**
E > F + G **E = 1.5 (F + G)**	**E = 1.5 (F + G)**
F = 2G	**F = 2G**

Next, assign the lowest-order alternative (in this case, **G**) a value of 1. Then back solve the system of equations to determine values for the set of alternatives. The result in this example is:

A = 17.5

B = 11.5

C > 5.5 and C < 6.5

D = 4.5

E = 4.5

F = 2

G = 1

Since the value for C is not exact, it can be approximated and assigned a weight of 6.

The sum of these weights is 47, so to normalize the values, divide each one by 47. The resulting numbers sum to 1 and give the weights. From A to G they are: 0.372, 0.245, 0.128, 0.096, 0.096, 0.043, and 0.020.

The paired comparison method can be used to find weights for the individual evaluation criteria and/or for the evaluation categories themselves. The table below shows the weights corresponding to individual evaluation criteria.

Table 3: Paired Comparison Weights shown on Evaluation Criteria Template

#	Evaluation Criteria	Description of How to Test the Criteria	Weight	<product P1 name>	<product P2 name>	<product P3 name>	<product P4 name>	<product P5 name>
1.0	Category 1							
1.1	Criteria A	-description-	0.372	0	0	0	0	0
1.2	Criteria B	-description-	0.245	0	0	0	0	0
1.3	Criteria C	-description-	0.128	0	0	0	0	0
1.4	Criteria D	-description-	0.096	0	0	0	0	0

Reference Comparison:

The Reference Comparison method is an alternative to the Paired Comparison and is a good alternative when calculating weights for 100+ criteria. Given a set of evaluation criteria, choose the evaluation criterion that is most important or significant in the set. Assign this criterion a value of 3. Using this as a reference, rank the remaining criteria as follows[4]:

- 3 = the criterion is as important as the "reference criterion"
- 2 = the criterion is slightly less important as the "reference criterion"
- 1 = the criterion is much less important than the "reference criterion"

Then, normalize these values so that they sum to 1.

For example, suppose values are assigned as follows:

A = 3

B = 3

[4] It is not necessary to use the range from 1 to 3. The range can be less constrained or more constrained as needed.

C = 2

D = 2

E = 3

F = 1

G = 2

The sum of these weights is 16, so to normalize the values, divide each one by 16. The resulting numbers sum to 1 and give the weights. From A to G they are: 0.1875, 0.1875, 0.125, 0.125, 0.1875, 0.0625, and 0.125.

The reference comparison method can be used to elicit weights for the individual evaluation criteria and/or for the evaluation categories themselves. The table below shows the weights corresponding to individual evaluation criteria.

Table 4: Reference Comparison Weights on Evaluation Criteria Template

#	Evaluation Criteria	Description of How to Test the Criteria	Weight	<product P1 name>	<product P2 name>	<product P3 name>	<product P4 name>	<product P5 name>
1.0	Category 1							
1.1	Criteria A	-description-	0.1875	0	0	0	0	0
1.2	Criteria B	-description-	0.1875	0	0	0	0	0
1.3	Criteria C	-description-	0.125	0	0	0	0	0
1.4	Criteria D	-description-	0.125	0	0	0	0	0

3.1.4 Computing the Overall Score for Each Product

Once the evaluation criteria, product scores, and evaluation weights have been determined, the *n*th additive utility function is used to compute the overall score of each product, where *n* is the number of evaluation criteria.

As an example, the additive utility function with two evaluation criteria, a_1 and a_2, is:

$$u(a_1, a_2) = w_1 u_1(a_1) + w_2 u_2(a_2)$$

The variables in the function are:

- u, represents the overall score of a product over two evaluation criteria, a_1 and a_2
- u_1 and u_2, scoring function(s) for criteria a_1 and a_2, respectively. For simplicity, teams can use the same scoring function for each criterion. The scoring function example from Section 3.1.2 demonstrated a constructed scale.

- w_1 and w_2, individual weights assigned to each criterion by a weight assessment method. The process of eliciting weights was described in Section 3.1.3.

Therefore in summary, MAU analysis provides evaluation teams with a consistent, fairly rigorous approach for scoring products in a technology evaluation. Teams must establish the evaluation criteria; determine a scheme for scoring products; and weight the relative importance of each evaluation criterion and category. The results are the collective efforts of evaluation teams, and are therefore likely to have some inter-subjective consistency. After each product has been evaluated and scored, the *n*th additive utility function gives the overall score (or utility) for each product and an overall product ranking.

3.2 Communication throughout the Evaluation Process

A successful evaluation requires effective communication between the evaluation team and the sponsor, stakeholders, subject matter experts, and vendors throughout the evaluation process. The team must understand what the problem is and what the solution is intended to accomplish.

During each phase, evaluation teams should conduct status updates with the sponsor and stakeholders and/or subject matter experts, either in writing or as a briefing, to discuss and solicit feedback on the following items:

- Evaluation goals and objectives
- Initial product assessments
- Additional products or currently deployed solutions within the sponsor's environment worth considering
- Considerations/requirements for the sponsor's environment
- Evaluation criteria and the test plan

In order to facilitate consistent, well-presented work during an evaluation that is recorded for later reference, Appendix B provides STEP briefing and document deliverable templates for each phase of the evaluation. In addition to ensuring good communication throughout the evaluation, the STEP templates also assist the team in drafting their final report.

3.3 Ensuring Evaluation Integrity

It is critical that MITRE teams perform evaluations that are recognized as comprehensive and fair. A fundamental requirement to achieving evaluation integrity is consistent documentation of test data and methodology for review by the sponsor, stakeholders, and vendors if questions arise. The STEP actions and tips (Chapters 4-6) provide guidance for ensuring evaluation integrity. These guidelines include:

- Verifying all product information for a Market Survey/Tool Selection with product vendors, and requesting written explanations (by email) as needed

- Following the rules and principles for establishing evaluation criteria, scoring products, and weighting criteria, as explained in Section 3.1
- Finalizing evaluation criteria, including associated weights, test procedures, and expected outcomes/guidelines for scoring before testing is begun.
- Highlighting product strengths and weaknesses as they are indicated in the overall evaluation scores. That is, the evaluation team must be careful not to call out product strengths and weaknesses arbitrarily in the final report without quantitative results and/or justification to back up the claims.
- Documenting the evaluation using STEP templates for consistency

3.4 Creating an Evaluation Timeline

Scheduling is an important part of the evaluation process in order to establish realistic timelines and expectations. The STEP workflow allows teams to identify the individual actions and estimate the time required to complete each one. Teams may wish to break larger actions into smaller segments to ensure that all of the evaluation work is well defined [13]. Teams must also work with their sponsor to determine the appropriate number of products to be tested with the time and resources available. Successful planning and timelines throughout the project will result in managing the work required for the evaluation.