A Formal Process for Evaluating COTS Software Products

A software product evaluation process grounded in mathematics and decision theory can effectively determine product quality and suitability with less risk and at lower cost than conventional methods.

As government agencies and businesses become more dependent on commercial off-the-shelf (COTS) software products to automate tasks, the ability to determine product quality and suitability has become increasingly important. Unfortunately, the lack of a standard, well-defined software product evaluation approach has resulted in many organizations making large investments in poor product choices.

An effective software product evaluation uses a formal process commensurate with the investment required for product acquisition and support to assess product quality and suitability prior to purchase. Anything less can only serve to justify a predetermined choice. Our requirements-driven COTS product evaluation process (RCPEP) ensures a quality outcome. We demonstrate its use in a case study that resulted in the selection of a product by the US Air Force for a large, component-based training management system. We believe that the successful application of this scientifically based method could mature into an industry standard.

**FORMAL EVALUATION PROCESS**

The formal RCPEP, shown in Figure 1, differs from a typical ad hoc product evaluation in a number of ways. To begin with, RCPEP relies on user-defined requirements instead of a short list of evaluator-generated criteria to determine product suitability and quality.

Also, an ad hoc process usually uses an arbitrarily chosen short list of candidate products. In contrast, RCPEP identifies every product that possibly addresses the requirements, conducts a trade study to narrow the list to serious candidates, and evaluates the remaining candidates using hands-on scenarios. For each product, evaluators determine how many requirements it satisfies as well as the quality and suitability of its implementation of these requirements using established criteria.

An effective evaluation process must apply decision theory, which makes it possible to quantify people’s choices. The simplest and most widely used calculations in decision theory rely on linear additive functions (weighted averages),¹,² which assume that the characteristics being evaluated are independent. Weighted averages are useful in analyzing both product requirements and evaluation criteria. Whereas ad hoc evaluations conduct little analysis of results using weighted averages, the RCPEP analysis is extensive.

Finally, an ad hoc process does not adhere to strict controls that ensure all candidate products receive identical evaluations with directly comparable results. RCPEP’s hands-on evaluation includes several controls:

- The same persons must evaluate each product.
- Each product must have the same configuration.
- Each product evaluation must use the same scenarios.
- Each product evaluation must use the same data.
- The evaluators must apply the same requirements and criteria to each product.

These controls ensure that the evaluation does not compromise the validity of the results.

**Trade study**

The first step in a trade study is establishing independent requirements for each product. Breaking down complex or interrelated requirements into simple components is essential because the validity of product comparisons depends on independent requirements at the same level of granularity. To determine how well a product meets these requirements, evaluation coordinators compose a questionnaire with one question per requirement and submit it to vendors of
potential candidate products. When possible, they also solicit current product users for more information on user satisfaction and vendor support.

**Composite requirements matrix.** Analysts examine the input from vendors and users and provide a consistent summary of results in a composite requirements matrix, illustrated in Figure 2. Matrix indicators identify the source of evaluation information and the breadth of requirement coverage.

Based on user inputs, analysts assign numerical weights ranging from 0 to 10 to each requirement to distinguish levels of importance. The more important a requirement, the higher its number. Because the weights are estimates, only single digits of precision are used. Analysts also assign numerical values to each requirement indicating the breadth of coverage for the candidate products. Requirements receive a 1 for full coverage, 1/2 for partial coverage, and 0 for no coverage.

Analysts multiply the numerical weight of each requirement by each candidate product’s requirement coverage score. Comparing each product’s sum to a perfect score—that is, the weighted representation of full requirements coverage—determines its requirements coverage ratio.

**Narrowing the field.** Products scoring less than half of the available points in the composite requirements matrix receive no further consideration. Those receiving substantially more than half of the possible points receive a recommendation for hands-on evaluation. If no product meets an acceptable level of requirements coverage, as determined by the customer organization, the trade study process must change to include better candidate products or different requirements. Failing this, automation of the required capabilities must rely on custom development rather than COTS products.

**Hands-on evaluation**

Analysts in cooperation with users create appropriate hands-on scenarios and criteria for rating the quality and suitability of each product with respect to requirement implementation. Evaluators selected by the using organization rate the products based on the criteria and product performance in the scenarios.

**Analysis requirements coverage matrices.** During the hands-on phase, evaluators refine the assessments of requirement coverage from the composite requirements matrix of the trade study. A consensus of the evaluators determines the overall coverage indicator for each requirement. If a large majority of evaluators do not agree on the level of coverage, partial coverage is assigned as the consensus value. Analysts again assign numerical values of 1 for full coverage, 1/2 for partial coverage, and 0 for no coverage.

---

**Figure 1. Requirements-driven COTS product evaluation process (RCPEP).** (a) The trade study results in a shortened list of candidate products that address a substantial percentage of user-defined requirements. (b) The hands-on evaluation quantifies the quality and suitability of implementation of these required features, providing an objective basis for product selection.
For each product, analysts multiply the numerical weight and assigned coverage value of each requirement and summarize the results in an analysis requirements coverage matrix. They then divide the sum by the total of the requirement weights and multiply by 10. This normalizes the results by putting the requirements ratings and all criteria ratings into the same 0 to 10 range.

**Evaluation criteria matrices.** The hands-on evaluation uses the scenarios to test how well each candidate product meets certain user requirements—for example, quality of window displays, value of online help, and adequacy of the analyses performed. The scenarios follow a step-by-step procedure to evaluate the same capabilities of each product. Evaluators rate applicable criteria to indicate how well each product performs. They use a 0 to 10 range to record these ratings in the cells of an evaluation criteria matrix for each scenario.

**Analysis criteria matrices.** Based on the weights of the related requirements, analysts assign weights to each criterion in the overall analysis criteria matrices for each scenario. They combine the evaluator responses from the evaluation criteria matrices into the corresponding analysis criteria matrix, as illustrated in Figure 3. To calculate the overall result \( R \), the analysts add the evaluators’ ratings for each criterion \( r_i \), multiply this sum by the corresponding weight for the criterion \( w_i \), and divide the result by the product of the sum of the weights \( W \) and number of evaluators \( N \). This formula, \( R = \frac{1}{W \times N} \sum w_i r_i \), produces a normalized overall rating in the familiar 0 to 10 range. Because all inputs to the calculations have one digit, the results are expressed using the same precision.

**Final analysis**

A final analysis follows the hands-on evaluation. Although the analysis requirements coverage matrices are used to more fully understand the requirements coverage for each product, the emphasis here is on the analysis of criteria ratings. The final analysis draws on several partial analyses of the collected criteria ratings.

The evaluators are divided into groups based on their different organizations, jobs, job locations, and other differences, and the evaluation criteria are similarly split into categories. Analysts place the results for each different grouping of evaluators or evaluation criteria into separate analysis criteria matrices and examine the differences and trends. These partial analyses determine the sensitivity of the overall results to the ratings of each individual and each group, and analysts can take group preferences into account when making final product recommendations.

**CASE STUDY**

We recently used RCPEP to evaluate a COTS product for the US Air Force’s Air Education and Training Command. As part of an overall strategy to imple-
ment a command-wide automated training management enterprise, AETC sought to insert a COTS product supporting training evaluation into its component-based Technical Training Management System (TTMS), which supports all technical training within the Air Force.

**Trade study**

We first conducted a trade study of 48 candidate products identified through a Commerce Business Daily request for information (the standard US government method for requesting information from potential contractors), an Internet search, and data gathered about other products in related projects. We contacted all of the products’ vendors, and 16 responded with completed questionnaires about how well their products met the TTMS requirements.

One major requirement category was overall product cost, which included a breakdown of costs as initial purchase, initial integration into system software infrastructure, initial user training, ongoing maintenance, ongoing user training, periodic upgrades, reintegration of upgrades into system software infrastructure, and user retraining for upgrades. AETC chose to examine this category of requirements separately rather than as part of the trade study.

Our analysts entered the appropriate requirements coverage indicators into the composite requirements matrix. They also contacted several of the products’ users and input data based on their comments into the matrix. An analysis of the composite requirements matrix resulted in the selection of three products for hands-on evaluation, which we refer to as P, Q, and R. All three were within the same general cost range.

**Hands-on evaluation**

We collaborated with a group of 17 evaluators to develop four scenarios and criteria to use in the TTMS evaluation sessions. The scenarios described survey, testing, and analysis functions for the products. We permitted the vendors to develop their own fifth scenario to showcase additional product capabilities. We released all of this information, as well as the data set to be used while running the evaluation scenarios, two weeks prior to the evaluation sessions so that the vendors would have ample time to prepare.

**Evaluators.** The evaluators represented five Air Force bases: Goodfellow, Lackland, Vandenberg, Keesler, and Sheppard. The first three groups each included one or two evaluators. For analysis purposes, we combined these three groups into one called GLV consisting of four evaluators, the same size as the Keesler group. Because the Sheppard group comprised nine evaluators, more than half of all evaluators, we combined the GLV and Keesler groups to create two approximately equal groupings. As the “Evaluator Group Composition” sidebar describes, this division provided an effective way to interpret assessments by people of different backgrounds and viewpoints.

The evaluators performed two types of roles. Seven evaluators focused on the development of tests for assessing student progress while in training. The remaining 10 concentrated on developing surveys for graduates and their supervisors to assess the effectiveness of the training courses.

**Preparation.** We set aside one calendar week for the hands-on evaluation process and allotted a full day to evaluate each of the final three products. We conducted a blind drawing to determine which day of the week a candidate product would be evaluated.

Because permitting each evaluator to have hands-on access to the product was impractical, we prepared a facility in which the evaluators could observe an analyst, who acted as their representative, executing each scenario while a vendor advised the analyst of the best way to accomplish each step. The evaluators sat at a table facing the vendor and analyst so they could observe what was happening without being directly visible to other evaluators.

**Evaluation sessions.** Each evaluation session began with the vendor providing a brief introduction to the product. The analyst then executed the first scenario while describing to the evaluators what was happening. The vendor provided continual guidance to the analyst to ensure proper use of the product.

At the end of the scenario, the evaluators each submitted a scenario description document and an evaluation criteria matrix complete with ratings. Each of the remaining scenarios repeated this pattern. A question-and-answer-period with the vendor followed the scenarios. The evaluators concluded the evaluation session by completing and submitting the requirements matrix.

It was clear on the first day of the evaluation sessions that the scenarios were too long and involved. Although accomplishing all of the scenarios’ steps was impractical, the evaluators dedicated about 10 hours...
to each product and accomplished all of their objectives. They unanimously indicated in a feedback session that the process was comprehensive and fair and would result in selecting the best product for the TTMS.

After completing the week-long evaluation process for the three candidate products, the evaluators reviewed their evaluation materials for consistency and made minor adjustments to their own ratings. The net effect of the changes, however, was negligible given the large number of evaluators.

Case study results

The AETC originally chose a scale of 1 (nice to have) to 4 (critically important) for the requirements and criteria weights. However, because TTMS evaluators assigned a weight of 4 in most cases, the weights were an insufficient discriminator in the evaluation analysis. To identify a potentially better discriminator, coordinators asked the evaluators to select their 10 most important requirements and criteria, which analysts used to assign enhanced weights for the most important requirements and criteria in some of the data analyses. The enhanced weights made no appreciable difference in the evaluation results.

Requirements coverage and criteria analyses

The hands-on evaluation considered four categories of product requirements: general product information, human-computer interface, product capabilities, and reports. Analysis of the requirements coverage revealed that product Q rated higher than product R overall, and product P was significantly lower than either of the others. Products Q and R outperformed P in every category.

Analysts performed partial analyses of the criteria ratings to determine the potential effect of the evaluators’ roles, the bases the evaluators represented, and the four product requirement categories on the ratings. Figure 4 shows the average criteria ratings for P, Q, and R by role, base, and category for all of the scenarios. The results repeated the same basic pattern as the requirements coverage analysis with individual exceptions in which R rated slightly higher than Q.

Final analysis

Controlling for a number of factors and considering many different views of the criteria ratings, RCPEP demonstrated that products Q and R were both acceptable because their ratings consistently exceeded a user-established threshold of 7. Product P was unacceptable because it could not meet that threshold. Product Q demonstrated slightly higher ratings than product R. Scenario 4 yielded the largest difference in ratings, with Q consistently rated higher than R. The AETC used the results of our analysis to select product Q.

The TTMS case study confirmed our expectation that the simplest form of decision analysis, weighted averages, would be effective in evaluating COTS products. RCPEP was successful because it relied on user-defined requirements, contained a sufficient number of evaluators to obtain valid quantitative results, and adhered to the controls.

We believe RCPEP will be effective for all forms of COTS product evaluations. However, many COTS product decisions do not warrant the expense involved in the application of the full RCPEP analysis, so we are working on ways to tailor the process to be com-
mensurate with the cost involved in the COTS product purchase without sacrificing its integrity.

Acknowledgments
This work was supported by the United States Air Force, Air Education and Training Command (HQ AETC/XPRI), under GSA Contract GS-35F-4885G.

References

Patricia K. Lawlis is part owner and senior consultant for c.j. kemp systems, an Ohio-based training and consulting firm. Her research interests include evaluating computer languages and software products, requirements analysis, and modeling of large, component-based training management systems. She received a PhD in software engineering from Arizona State University. Lawlis is a member of the IEEE Computer Society and the ACM. Contact her at lawlis@aol.com.

Kathryn E. Mark is part owner, engineering assistant, and evaluation coordinator for c.j. kemp systems. Her research interests include evaluating component products for large, component-based training management systems. Contact her at KEMark@aol.com.

Deborah A. Thomas is a principal analyst and member of the Corporate Systems and Software Engineering Process Improvement Committee for Jacobs Engineering/Sverdrup Technology. Her research interests include requirements analysis and modeling of large, component-based training management systems. Thomas received an MS in software engineering and computer science from the University of West Florida. She is a member of the ICCP and the AOC. Contact her at ThomasDAT@aol.com.

Terry Courtheyn is a principal analyst for Jacobs Engineering/Sverdrup Technology. His research interests include requirements analysis and modeling of large, component-based training management systems. He received an MS in electrical engineering from the Air Force Institute of Technology. Courtheyn is a member of the AFCE and the AOC. Contact him at terry.courtheyn@jacobs.com.

Introducing the IEEE Computer Society Career Service Center

Advance your career
Search for jobs
Post a resume
List a job opportunity
Post your company’s profile
Link to career services

computer.org/careers/