MTR070061R1

MITRE TECHNICAL REPORT

Defining Useful Technology Evaluations

13 September 2007

Paul M. Herceg

(Revised from 17 October 2006)

Project No.: 0707G01Z-HB

MITRE

Center for Integrated Intelligence Systems
McLean, Virginia

MITRE Department
and Project Approval:

____
Michael E. Shea, Project Leader

Abstract

Information technology executives and managers at MITRE sponsors are the decision-makers who direct the selection and adoption of new technologies. Nonetheless, they often lack a systematic method for assessing technology value. Although best practices are well documented in academic literature, decision-makers may unknowingly fall victim to or use haphazard approaches that fail to meet business needs and bypass more effective technologies. This paper explains to executives and managers why they need a systematic approach to technology evaluation. Then, it distills for them the evaluation methods and technology investigation processes proposed by experts in empirical software engineering. This paper has not aimed to add novel ideas to the academic discussion of technology evaluation. Rather, the intent here is to allow MITRE sponsors to access, understand, and apply these concepts.

Keywords: Technology evaluation, technology assessment, technology insertion, industry

Table of Contents

Introduction

In the information technology (IT) organization, executives and managers are the technology decision-makers and are responsible for assessing the degree to which each available technology fulfills business needs. Often, these leaders initiate projects to assess technology value, called *technology evaluations* or *technology assessments*. These projects generate results that inform the leaders' technology acquisition decisions. Ideally, evaluations should contain a mix of priority user needs and promising technologies. Achieving this mix within cost and schedule constraints is difficult without a process. Business needs and budgets change, and executives can find themselves deciding between (a) completing an evaluation effort in the face of higher priorities, or (b) ceasing an evaluation activity, accepting the sunk costs, and adopting a technology without ample justification. A well-defined process can help leaders avoid this undesirable situation, obtain the guidance needed, and demonstrate fiscal responsibility.

In open literature, there is a wealth of guidance on technology evaluations and technology investigation processes. This knowledge tends to remain unavailable and inaccessible to IT executives and managers at MITRE sponsor sites because it is distributed across academic publications addressed to audiences outside the government. This paper transforms this expert guidance into a form that IT leaders can comfortably comprehend and apply. This paper does not aim to add new ideas to the body of knowledge on technology evaluations, but rather distills extant knowledge to make it accessible, available, and sharable among MITRE sponsors. It describes the importance of evaluations, the types of evaluations, and a process that can be applied to define high-impact industrial evaluations. Although the concepts in this paper can be applied to other domains, this paper was tailored for organizations interested in document and media processing technologies.

The Technology Decision-Maker's Challenge

Requests to adopt new technology come from executives, peer staff, vendors, and universities, among others. In the absence of quantitative results or expert qualitative assessment, organization decision-makers are often left to select technologies based on intuition. IT and software manufacturing executives often find themselves asking the following types of questions, and they grope for the answers:

- How can we use technology X to solve our problem?
- This technology looks good. Should we buy this one instead of the one that's slated for procurement?
- Which technologies should we include in the prototype?
- Is this technology a match for my enterprise?
- Is this vendor's new algorithm really better than the state-of-the-practice algorithms?
- How does this vendor's tool compare against the other vendor tools in this market space?
- I don't have funding for the evaluation, but we've got to know if we should go with this new technology or stay with our incumbent technology.
- We have an operation to run. We don't have time for an evaluation. But how can I make a procurement decision without validating that the technology works?
- We need an evaluation to quantify benefit, but that will take too long. But how do we advise management in the allotted 30 days without an evaluation?

- We need an evaluation to understand which technology to use for the prototype, but that will take too long. But how do we ensure we're integrating best-of-breed components?
- Which new technologies should we focus evaluations on?
- How do we know that these technology evaluations will have high impact on our organization?
- Can we even afford evaluations?
- Do the near-term benefits of this new technology justify the huge cost of moving our enterprise to use it?!

There is at least one field of academic literature that is clearly addressing these questions, and in a manner that can be applied to industry.

Experts in the field of empirical software engineering explain that a variety of validation approaches exist for making better informed technology decisions; in this paper, these are referred to as *evaluations*. Among the opinions of experts in empirical software engineering are the following:

- It is quite obvious…that technology evaluation is not for free. It does take time and effort to evaluate technology, but…we cannot just change due to trends [Wohlin96, p. 5].
- Often new software technology is touted as the next "silver bullet" to be adopted, only to fail and disappear within a short period. New technologies are often adopted without any convincing evidence that they will be effective, yet other technologies are ignored despite the published data that they will be useful [Zelkowitz03, p. 230].
- [There are] tremendous latency times between the invention of new technologies and their wide-spread acceptance…ranging between 11 and 23 years…A main reason for these long latency times lies in the uncertainties about the effect of new technologies on one's own (target) environment [Houdek03, p. 161].
- Technology features alone, however, are insufficient for us to understand added value; for this, we must understand how these features will be used and what benefits will accrue from their use…. Situating a technology in a technology marketplace and in a problem domain gives us a basis for identifying feature deltas [Brown96, p. 42].

Dybå repeats the same motto of those in this field who recommend that decision-makers base their technology selection on evidence: "Software companies are often under pressure to adopt immature technologies because of market and management pressures. We suggest that practitioners consider evidence-based software engineering as a mechanism to support and improve their technology adoption decisions" [Dybå05, p. 59].

Evaluation should involve identifying the features that discriminate one technology from another. Brown states that "a technology's potential impact is best understood in terms of its feature delta…to understand its added value, you must focus on its distinctive features in relation to other technologies" [Brown96, p. 41]. For emerging technologies, such as in the area of human language technology, accuracy is often the distinctive feature. For technologies that are ubiquitous in the commercial space, a larger breadth of features must be considered.

The Cost of No-Validation Versus the Reward of Evaluations

Information infrastructure integration projects using best-of-breed technologies can vary from millions to hundreds of millions of dollars; poor technology decisions multiply these costs to the immediate project or program and have hidden costs to production operations. Dybå established that a lack of evaluation leads to these poor technology decisions.

> Software managers and practitioners often must make decisions about what technologies to employ on their projects. They might be aware of problems with their current… practices…and want to resolve them. Or, they might have read about a new technology and want to take advantage of its promised benefits. However, practitioners can have difficulty making informed decisions about whether to adopt a new technology because there's little objective evidence to confirm its suitability, limits, qualities, costs, and inherent risks. This can lead to poor decisions about technology adoption. [Dybå05, p. 58]

The costs of poorly matched or poorly performing technologies are obvious for the domain of interest to this paper. First, there is the opportunity cost of not finding critical information of interest. Second, there is the sunk cost of integrating the under-performing technology. Third, years of unplanned effort and cost are expended to reverse the integration of the poorly performing technology. Finally, additional unplanned effort and cost are needed to find and insert a technology that performs well.

Evaluations aid the adoption of technologies that provide compelling benefits. However, conducting evaluations alone is insufficient. Proper planning is required to ensure responsible and ethical use of funding. Not thinking through the objectives and methodology for evaluations leads to a lack of benefit to production environments and a reduction in funding for future evaluations. The lack of benefit to production can be attributed to narrowly focused or misfocused evaluations that have impact for a single use case in a single domain.

On the other hand, a technology investigation process that includes proper planning leads to evaluations that have broad applicability and impact. This is a responsible use of evaluation funding and leads to a positive impact on production operations and expanded evaluation resources. Furthermore, it makes the most of limited evaluation funds. Proper planning includes a clear technology insertion mission, investigating evaluation methodologies and data, use case elicitation, technology scouting, collaboration with partner organizations, technology candidate and evaluation candidate determination and prioritization, and maintaining an experience base. Each of these is described later in this paper.

Defining Useful Evaluations Involves Asking the Right Questions

Before embarking on a description of the wide variety of evaluation methods and the technology investigation process, it is helpful to get a better grasp of what constitutes an evaluation. An evaluation can be viewed as a process, with clear starting and ending points. At an abstract level, the process comprises the following steps:

1. Ask a question about a technology or technologies
2. Obtain relevant data
3. Interpret the data
4. Answer the question

For example, think of a software company executive who asks which of a number of related technologies should be adopted in his or her organization (Step 1). Typically, an evaluation effort provides the answer (Step 4). Upon receiving this answer, the executive decides on a technology direction that makes sense consistent with budget, schedule, and mission. The interim steps—obtaining and interpreting the relevant data (Steps 2 and 3)—are often referred to as the evaluation *method* or *methodology* and are discussed in the next section.

Asking the right question is not as easy as it seems. Dybå explained that evidence-based software engineering involves refining initial questions [Dybå05, p.60]. Often in industry, the first question posed is not the question that will contribute to an informed technology decision. For example, the question "How can we use technology X to solve our problem?" may be better stated "How does technology X compare among the available alternatives to solve our problem?" And this question then gets decomposed into "What are the available alternatives?", "How do the available alternatives rank?", and "Among the three most promising alternatives, which is the best fit for us?" So, IT executives need to deliberate over the questions they ask to ensure that the answer they receive will help them make effective, responsible technology decisions. This questioning process is the focus of this paper. A systematic approach is needed to formulate the right technology questions. And a carefully thought-out planning process can ensure that an organization defines evaluations in such a way that the results are truly useful.

Evaluation Methods

An abstract understanding of evaluations makes it possible to discuss various evaluation methods. Evaluation methods involve obtaining data about a technology or set of technologies and interpreting that data. Such a method can be as simple as a single person collecting qualitative data via a literature search and analyzing the information, or as complex as a collecting quantitative data via production system instrumentation and analyzing that data. The former would provide a high-level picture of a technology's effectiveness, whereas the latter could be used to actually validate effectiveness.[1]

A survey of papers on empirical software engineering reveals a number of evaluation methods. Zelkowitz provided the most comprehensive and helpful catalog of evaluation methods [Zelkowitz03, pp. 236, 243]. They identified 14 methods used in academia and 15 methods used in industry, defined each, and discussed in detail the relative merits of each. Wohlin and Houdek presented similar catalogs when they presented their technology adoption frameworks [Wohlin96, p. 6] [Houdek03, p. 145]. Wohlin presented the most understandable grouping of evaluation methods by relating them according to the evaluation environment required. According to him, as environment progresses toward production, several factors increase, including the cost, the similarity to the production environment, and the quality of the evaluation result (i.e., the quality of the evidence to affect technology decisions). Typically, staffing and schedule also increase. A simplified list of industrial evaluation methods, combining the methods from Zelkowitz, Wohlin, Houdek, is shown in Table 1.[2] This table also reflects the notion of

[1] Researchers use evaluation to validate or refute the usefulness of a technology they are developing.
[2] See additional discussion of specific evaluation methods in [Basili86, pp. 735-740], [Brown96, p. 43], and [Wohlin00, pp. 10-14].

increasing factors. Weak methods that should be avoided are listed in Table 2; such methods require subjective persuasion to convince people that a technology has benefits.

Both Wohlin and Houdek discuss the cost progression as the evaluation environment changes from desktop, to laboratory, to a production environment (see Table 1). In other words, desk methods may be performed by a single staff person, whereas laboratory methods will require a specialized team and a can require a moderate amount of equipment. The most elaborate methods are those related to production and because they involve deployed systems, they require significant investment.

Often decision-maker technology questions can be answered by a desktop evaluation, and this simpler approach avoids the cost of more resource-intensive methods. However, the simple approach is not always the best approach. How the evaluation results will be used is an essential factor in determining the rigor needed in evaluation. For example, technology decisions for emerging technologies often necessitate more advanced forms of evaluation.

Table 1. Industrial evaluation methods

Evaluation Environments	**Evaluation Methods**	**Factors: Cost, Staffing, Schedule, Similarity to Production Environment, Quality of Evidence to Affect Technology Decisions**
Desk	Expert Discussion and Opinion Feature Study/Vendor Literature Study Academic Literature Study Internal Field Study External Field Study Historical Data Study Theoretical Analysis	Low
Laboratory	Feasibility Test Experiment with Non-Random Data Formal Experiment	Medium
Production	Pilot Project Project Case Study Replicated Project Case Study	High

Adapted from Zelkowitz03; Wohlin96; and Houdek03.

Table 2. Weak evaluation methods to avoid

- No evaluation
- Non-expert opinion (e.g., manager's opinion)
- Vendor opinion
- Demonstration

The degree to which the evaluation environment reflects the envisaged or existing production environment is a key consideration when selecting an evaluation method. Wohlin describes this as *similarity* to the production environment [Wohlin96, p. 5], whereas Houdek describes it as the *distance* from the target environment [Houdek03, p. 135]. Nevertheless, the more closely the evaluation environment matches the envisaged or existing production environment, the more confidence the decision-maker will have in the results and the more the evaluation will contribute to a sound technology decision.

Depending on the quality of evaluation result required, accepting a moderate similarity may result in actionable information and a more effective use of evaluation funding than narrow, high-similarity evaluations. For example, if generalized evaluation results can apply to multiple production environments, it may be preferable to conduct a single generalized evaluation and apply the evaluation results to the multiple environments.

A Technology Investigation Process for Defining Useful Evaluations

A well-executed technology investigation process is driven by a clear technology insertion mission (see Figure 1) and includes planning steps to define useful evaluations. These steps position an organization to answer the questions "Which new technologies should we focus evaluations on?" and "How do we know that these technology evaluations will have high impact on our organization?" Processes for defining useful evaluations have been well documented in academic literature. Both [Wohlin96] and [Houdek03] present complete and customizable process frameworks that managers can adapt to their individual industry needs.[3] Figure 2 depicts these steps as a fluid, recursive process.

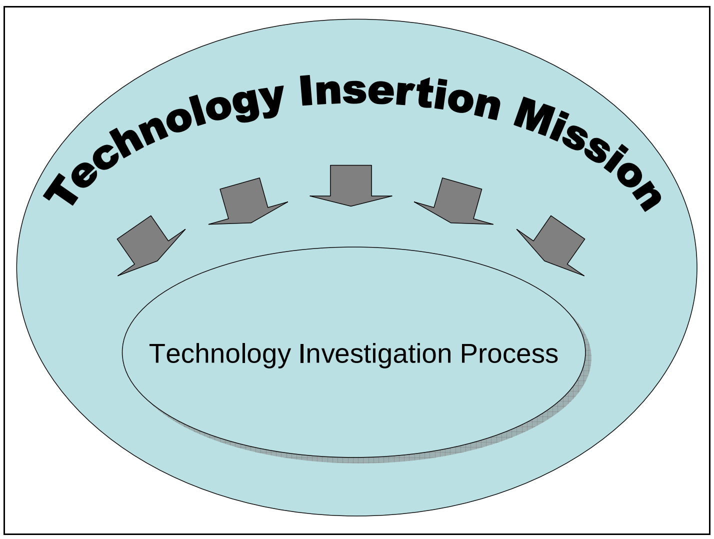

Figure 1. A technology insertion mission drives a technology investigation process.

The following steps, adapted from the [Wohlin96] and [Houdek03] frameworks, elaborate on the graphic:

- *Investigating evaluation methodologies and data* is vital in order to reduce the delay for designing and implementing evaluations. This is especially true for emerging technologies where evaluation methodologies are forming and evolving. For human language technology evaluations, corpus development can eclipse other aspects of an evaluation in terms of schedule and resources. Therefore, ongoing procurement of data that could be useful for evaluations is critical. Also, significant effort may be required to develop evaluation tools (e.g., software) if they are not already available.

[3] [Brown96] and [Basili86] offer similar, yet more involved framework descriptions..

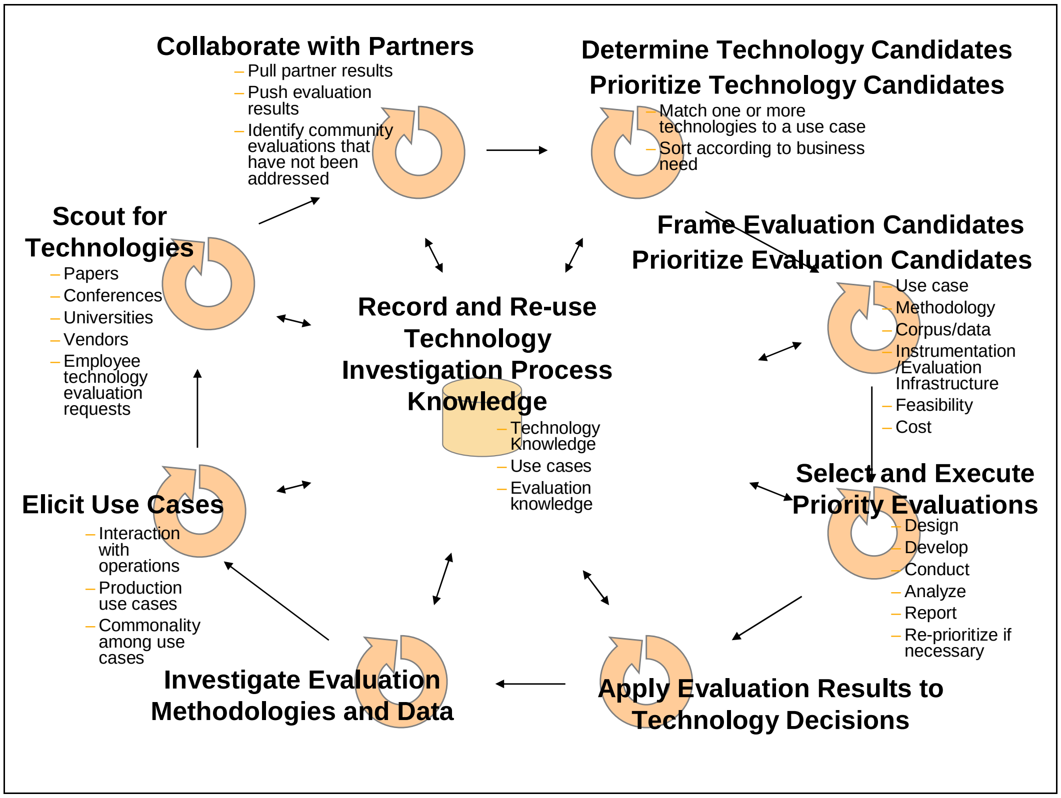

Figure 2. A technology investigation process.

- *Eliciting use cases* situates technologies in a problem space.[4] This step involves interaction with operations staff to identify the most pressing needs facing the organization. Conducting this step allows commonality among use cases to be identified, and enables the design of generalized evaluations that will have a wide-reaching impact on multiple target environments.
- *Scouting for technologies* exploits all sources where promising technologies may be found: reviewing papers and conferences, monitoring the advances of universities and vendors, and receiving technology evaluation requests from organization employees.
- *Collaborating with partner organizations* enables an organization to monitor the research and evaluations being performed elsewhere, to identify gaps (i.e., which critical evaluations have not yet been conducted), to pull and push evaluation reports from and to partners, and to influence vendor/university solutions.
- *Determining technology candidates* (i.e., technology screening) is the process of reducing an overwhelming abundance of technology options by matching them to organization use cases.
- *Prioritizing technology candidates* focuses on sorting each of these technology and use case matches so that the most promising matches are listed first.
- *Framing evaluation candidates* involves transforming priority technology candidates into outlines that describe potential evaluations. This exercise involves preliminary scoping of the evaluations, including identification of the evaluation use case, related production

[4] "Use cases, stated simply, allow description of sequences of events that, taken together, lead to a system doing something useful" [Bittner03, p. xvi].

use cases, methodology to be used, corpus/data to be used, evaluation infrastructure (i.e., instrumentation) to be developed[5], feasibility of the evaluation, and cost.

- *Prioritizing evaluation candidates* systematically sorts evaluation candidates along with the list of in-progress evaluations so that the high-impact evaluations are listed first. Availability of funding; time; skilled staff; appropriate facilities (e.g., a lab); and existing evaluation infrastructure, methodology, and data are all factors affecting the relative position of evaluation candidates. These factors should be estimated. Initially, an evaluation candidate can require far too many resources; however such candidates can be decomposed into component evaluations and then be re-proposed. Also, the anticipated scope of impact (i.e., benefit) for an evaluation candidate can be a significant and useful factor in assessing relative merit (see Figure 3). Houdek proposed that evaluation candidates should be ranked based on the following factors: anticipated quality of the evaluation outcome (i.e., correlated to the evaluation environment's similarity to production), the resources required, and the expected impact to the organization [Houdek03, pp. 142-146]. Some in-progress evaluation projects may be labeled low priority, re-positioned in the list, and considered for close-down; however this should happen rarely if an ongoing technology investigation process has been employed. This repositioning is vital to ensure that resources remain focused on high-impact evaluations.
- *Selecting and executing priority evaluations* involves initiating projects for the priority evaluation candidates. The researcher or evaluator designs the evaluation, including use case, methodology, corpus/data, evaluation infrastructure, outcome format, feasibility of evaluation, and cost of evaluation. Execution involves collecting, analyzing, and reporting the data. Upon completion, results should be recorded and shared with the organization and, if appropriate, community partners.
- *Recording and re-using technology investigation process knowledge* employs a central repository called an *experience base* [Basili94]. Both Houdek and Wohlin advocate this approach. Over time, an organization gains a significant amount of knowledge about technology, use cases, and evaluations that is vital to reviewing future evaluation candidates and requests. The database can be online or merely in a filing cabinet. All of the documentation related to the planning process and technology evaluations is recorded in the database, including evaluation results and recommendations. Organization use cases, papers, vendor literature, and evaluation reports from partners should be stored as well. These data are useful during the review of new evaluation requests. For example, if an evaluation being requested has already been conducted, the organization can retrieve the evaluation results and share them with the requestor. The database also serves as a resource for custodians of production systems; upon querying the database they may find technologies that provide solutions to operational problems.
- Finally, the cycle ends with achieving the overall process objective: *Apply evaluation results to technology decisions*. In this step, the decision-maker seeks to understand the evaluation results and uses this information as a factor in plotting technology direction.

A systematic approach that follows these steps—in a way tailored to the specific needs of the industry—is obligatory if the technology decision-maker is to ask the right questions about a

[5] For desktop evaluations, this may be as simple as a table in which to collect data and record qualitative judgments.

technology or set of technologies. This method of technology investigation ensures that the organization conducts only the evaluations that have the highest impact on technology decisions.

Sufficient personnel are required to realize this process. However, the staff does not have to be large. An evaluation-focused staff may begin as two or three dedicated people and may evolve into a larger department.

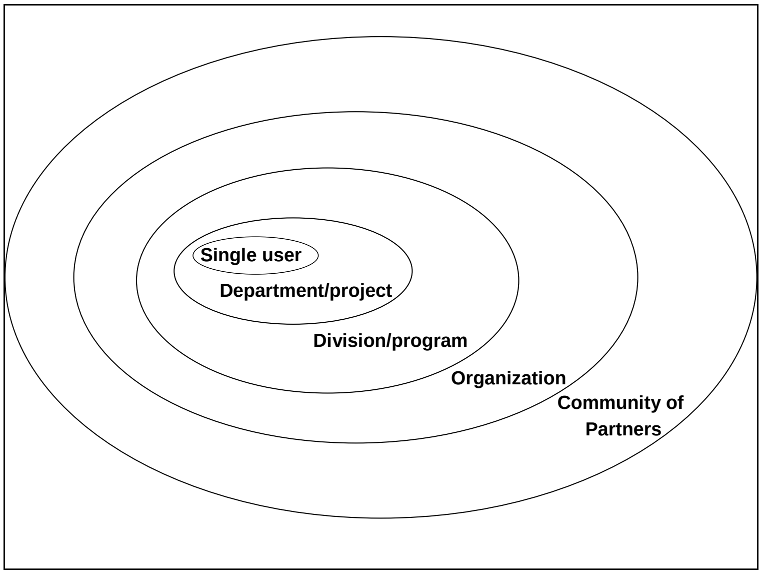

Figure 3. Evaluation candidate scope of impact.

Capturing and Reviewing Employee Evaluation Requests

Although technology scouting should be implemented by assigned staff, evaluation requests will inevitably come from multiple sources; therefore, it is necessary to have a mechanism to capture and regulate these requests. If allowed, organization employees will generate more evaluation requests than the staff assigned to implement the process. A simple form can help, and a form-based process allows each request to be reviewed, compared against the knowledge database, and prioritized. The following components are essential elements of an evaluation request and can be the start of a form for staff to complete; these are adapted from the structure in [Houdek03, p. 141]:

- Question: The question about a technology or technology area
- Technology: A description of the technology or technology area
- Production Use Case: A description of how the technology could or should be employed
- Evaluation Use Case: The specific use case against which the technology should be evaluated
- Recommended evaluation method: a description of the evaluation design, constraints, measures (qualitative or quantitative)
- Outcome format: Expectations for how the result should be reported (e.g., an executive summary of the results, a description of the technical results, or a comprehensive paper including the methodology and results)
- Impact: A description of the way the results of the evaluation are to be applied and a list of the specific decisions that are to be affected by the evaluation results

Evaluation requests have an array of benefits. New requests may point to the need to re-evaluate technologies for which evaluation results have become dated. A multiplicity of similar evaluation requests from different divisions can reveal that a technology area is important to the organization and that the formulation of a generalized evaluation may satisfy the multiple requests. Evaluation requests may reveal where systems engineering assistance is needed in the organization. Closer review of an evaluation request may reveal that the requestor needs:

- An awareness of technologies that better address their business need or use case
- An awareness of similar use cases and motivation to collaborate with other divisions to re-propose an evaluation with greater impact
- An awareness of evaluations already performed that can be accessed in the organization's tracked knowledge
- Assistance with system design
- Help understanding the true business need

In these cases an evaluation is not the best solution for the requestor or the organization; however dialog with the requestor will elicit the true needs, against which the proper skill-sets can be applied.

Conclusion

This paper distilled academic discussions of technology evaluations into a form that can be understood and implemented by IT executives and managers. Formulating the right technology question, understanding the array of evaluation methods, and implementing a technology investigation process are keys to successfully defining high-impact industrial evaluations. Investigation of evaluation methodologies, use case elicitation, technology scouting, collaboration, technology candidate and evaluation candidate determination and prioritization, and maintaining an experience base are all vital to realizing the most useful technology evaluations.

References

[Basili86] Basili, V. R., Selby, R. W., & Hutchens, D. H. (1986). Experimentation in software engineering. *IEEE Transactions on Software Engineering, 12*(7), 733-743. Retrieved September 13, 2007, from http://www.cs.umd.edu/~basili/publications/journals/J29.pdf.

[Basili94] Basili, V., Caldiera, G., & Rombach, H.D. (1994). Experience factory. In J. J. Marciniak (Ed.), *Encyclopedia of Software Engineering* (Vol. 1, pp. 469-476). New York: John Wiley & Sons. Retrieved July 27, 2006, from http://www.cs.umd.edu/~basili/publications/technical/T85.pdf.

[Bittner03] Bittner, K., & Spence, I. (2003). *Use Case Modeling*. Boston, MA: Addison-Wesley.

[Brown96] Brown, A. W., & Wallnau, K. C. (1996). A framework for evaluating software technology. *IEEE Software, 13*(5), 39-49. Retrieved July 27, 2006, from http://ieeexplore.ieee.org/iel5/52/11391/00536457.pdf (IEEE publications online).

[Dybå05] Dybå, T., Kitchenham, B. A., & Jørgensen, M. (2005). Evidence-based software engineering for practitioners. *IEEE Software*, *22*(1), 58-65. Retrieved July 27, 2006, from http://ieeexplore.ieee.org/iel5/52/30054/01377125.pdf (IEEE publications online).

[Houdek03] Houdek, F. (2003). External experiments—A workable paradigm for collaboration between industry and academia. In N. Juristo & A. M. Moreno (Eds.), *Lecture Notes on Empirical Software Engineering* (pp. 133-166). River Edge, NJ: World Scientific.

[Wohlin96] Wohlin, C., Gustavsson, A., Höst, M., & Mattsson, C. (1996). A framework for technology introduction in software organizations. In *Proceedings Software Process Improvement Conference* (pp. 167-176). Brighton, UK.

[Wohlin00] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M., Regnell, B., & Wesslén, A. (2000). *Experimentation in Software Engineering*. Boston; Kluwer Academic Publishers.

[Zelkowitz03] Zelkowitz, M. V., Wallace, D. R., & Binkley, D. W. (2003). Experimental validation of new software technology. In N. Juristo & A. M. Moreno (Eds.), *Lecture Notes on Empirical Software Engineering* (pp. 229-263). River Edge, NJ: World Scientific.

Additional Resources

Segal, J. (2003). The nature of evidence in empirical software engineering. In *Proceedings of the Eleventh Annual International Workshop on Software Technology and Engineering Practice (STEP'03)* (pp. 40-47). IEEE. Retrieved July 27, 2006, from http://ieeexplore.ieee.org/iel5/9448/29998/01372131.pdf?arnumber=1372131 (IEEE publications online).

Fenton, N., Pfleeger, S. L., & Glass, R. L. (1994). Science and substance: a challenge to software engineers. *IEEE Software, 11*(4), 86-95. Retrieved July 27, 2006, from http://ieeexplore.ieee.org/iel1/52/7423/00300094.pdf?arnumber=300094 (IEEE publications online).

Zelkowitz, M. V. & D. R. Wallace (1998). Validating the benefit of new software technology, *Software Quality Practitioner* 1(1). Retrieved July 27, 2006, from http://www.cs.umd.edu/~mvz/pub/sqp.pdf.