DESMET: a methodology for evaluating software engineering methods and tools

by Barbara Kitchenham, Stephen Linkman and David Law

DESMET was a DTI-backed project with the goal of developing and validating a methodology for evaluating software engineering methods and tools. The project identified nine methods of evaluation and a set of criteria to help evaluators select an appropriate method. Detailed guidelines were developed for three important evaluation methods: formal experiments, quantitative case studies and feature analysis evaluations. This article describes the way the DESMET project used the DESMET methodology both to evaluate the methodology itself and to provide direct assistance to the commercial organisations using it.

Software engineers split into two rather separate communities: those who build and maintain software systems, and those who devise methods and tools that they would like the former group to use to build their software systems. To the disappointment of the tool and method developers, adoption of new methods and tools by builders of software is slow. To the disappointment of the builders, new methods and tools continue to proliferate without supporting evidence as to their benefits over existing approaches. Indeed readers of technical journals might be forgiven for believing that software engineering is *only* concerned with the development of new methods, and the tools to support those methods.

Given this obsession with developing methods and tools, and the difficulty method/tool developers have getting other engineers to use them, it is surprising how little effort is directed towards evaluating methods and tools. In the late 1980s the DTI made a concerted effort to redress this oversight by part funding three research projects investigating the issue of evaluation. These were:

- the SMARTIE project which concentrated on the issue of evaluating software engineering standards[1]
- the META project which looked at the problem of evaluating design methods*
- the DESMET project which attempted to develop and validate a scientifically-based and practical approach to evaluation in the field of software engineering.

The DESMET project is the subject of this article. The project identified nine methods of evaluation and a set of criteria to help evaluators select an appropriate method. Detailed guidelines were developed for three important evaluation methods: formal experiments, quantitative case studies and feature analysis exercises. The DESMET evaluation methodology is intended to support individuals performing the following roles:

- a method or tool vendor seeking to demonstrate the advantages of his/her product
- a software process engineer responsible for assessing a proposed process change
- an academic researcher developing or investigating a new method
- a member of an ESSI consortium undertaking an evaluation of a new method in an industrial context.

The details of the approach are being published in a series of articles in *SIGSOFT Notes* by Kitchenham[2] and Pfleeger[3] and we will present only an introduction here.

*For details contact Prof. Bob Cole, TRICAT Ltd., 45 Victoria Road, Lenzie, Glasgow G66 5AP.

Readers may contact the authors at the University of Keele if they are unable to obtain the above articles.

Evaluations are always difficult; however, one of the most difficult items to evaluate is a methodology, such as the DESMET evaluation methodology itself. It is therefore of particular interest that DESMET applied the methodology it had developed to its own validation, providing a ready-made study of the problems and the pitfalls of undertaking such types of evaluation.

The DESMET project

The goals of the DESMET project were twofold:

1 to develop a methodology for evaluating software engineering methods and tools
2 to validate that methodology.

The project team comprised the National Computing Centre, the University of North London, GEC-Marconi and BNR-Europe. NCC was the lead partner in the project and, with the University of North London, took responsibility for developing the DESMET methodology. GEC-Marconi and BNR-Europe were responsible for evaluating the methodology.

Research strategy

The project started by reviewing the literature on software method/tool evaluation and reviewing past evaluation exercises.[4] In addition, the project noted that other disciplines faced similar evaluation problems: medicine (evaluation of medical treatments) and education (curriculum development). We found that a brief review of these disciplines proved to be a useful starting point for DESMET.

After our initial review, we identified a number of problems with the way in which evaluation exercises had been performed. These included:

1 *Difficulties defining appropriate controls:* In formal software experiments aimed at evaluating software development methods, the 'control' situation is often defined as 'the lack of a method', so effects of a new method are confounded with the effect of using any method. In industrial case studies the 'control' is the status quo and may be poorly defined or context dependent, making results difficult to generalise.
2 *Scaling problems:* Formal experiments are usually performed on relatively small software applications or tasks. Case studies can be performed on 'real' projects, but cost prohibits adequate replication.
3 *Evaluation costs:* Evaluations are expensive and there are no mandatory requirements on method/tool developers or vendors to validate their methods/tools.
4 *Difficulty defining objective benefits for certain types of technology:* For example many management tools/methods are risk reduction technologies not cost benefit technologies.
5 *Immature measures:* There are few universally agreed measures of software quality and productivity. This means that it is often difficult to replicate experimental results.
6 *Difficulties controlling or eliminating confounding factors:* Major evaluation problems result from variability in staff capability, learning-curve effects and expectation effects.

The most significant result of this work, however, was the realisation that there was no one method of evaluation that was always best. There were many different evaluation methods each of which was appropriate in different circumstances. Thus, our first research goal was to establish a taxonomy of evaluation methods and identify the conditions that would favour using one method rather than another.

Evaluation methods and selection criteria

The DESMET evaluation methodology separates evaluation exercises into two main types:

- *quantitative* evaluations aimed at establishing measurable effects of using a method/tool
- *qualitative* evaluations aimed at establishing method/tool appropriateness, i.e. how well a method/tool fits the needs and culture of an organisation.

Measurable effects are usually based on observed changes in production, rework or maintenance time or costs. DESMET refers to this type of evaluation as a *quantitative* or *objective* evaluation. Quantitative evaluations are based on identifying the benefits you expect a new method or tool to deliver in measurable terms and collecting data to determine whether the expected benefits are actually delivered.

The appropriateness of a method/tool is usually assessed in terms of the features provided by the method/tool, the characteristics of its supplier and its training requirements. The specific features and characteristics included in the evaluation are based on the requirements of the user population and any organisational procurement standards. Evaluators assess the extent to which the method/tool provides the required features in a usable and effective manner based (usually) on personal opinion. DESMET refers to this type of evaluation as *feature analysis* and identifies such an evaluation as a *qualitative* or *subjective* evaluation.

Some methods involve both a subjective and an objective element. DESMET calls these *hybrid* methods.

In addition to the separation between quantitative, qualitative and hybrid evaluations, there is another dimension to an evaluation: the way in which the evaluation is organised. DESMET has identified three rather different ways of organising an evaluation exercise:

Table 1 Evaluation method selection

Evaluation method	Conditions favouring method
quantitative experiments	Benefits clearly quantifiable. Staff available for taking part in experiment (i.e. perform 'non-productive' work). Method/tool related to a single task/activity. Benefits directly measurable from task output. Relatively small learning time Desire to make context independent method/tool assessments.
quantitative case studies	Benefits quantifiable on a single project. Benefits quantifiable prior to product retirement. Stable development procedures. Staff with measurement experience. Timescales for evaluation commensurate with the elapsed time of your normal size projects.
quantitative surveys	Benefits not quantifiable on a single project. Existing database of project achievements including productivity, quality, tool/method data. Projects with experience of using the method/tool.
feature analysis—screening mode	Large number of methods/tools to assess. Short timescales for evaluation exercise.
feature analysis—case study	Benefits difficult to quantify. Benefits observable on a single project. Stable development procedures. Tool/method user population limited. Timescales for evaluation commensurate with the elapsed time of your normal size projects.
feature analysis—experiment	Benefits difficult to quantify. Benefits directly observable from task output. Relatively small learning time. Tool/method user population very varied.
feature analysis—survey	Benefits difficult to quantify. Tool/method user population very varied. Benefits not observable on a single project. Projects with experience of using the method/tool, or projects prepared to learn about the method/tool.
qualitative effects analysis	Availability of expert opinion assessments of methods/tools. Lack of stable development procedures. Requirement to mix and match methods/tools. Interest in evaluation of generic methods/tools.
benchmarking	Method/tool machine-intensive rather than human-intensive. Outputs of method/tool able to be ranked in terms of some 'goodness' criteria.

- as *a formal experiment* where many subjects (i.e. software engineers) are asked to perform a task (or variety of tasks) using the different methods/tools under investigation. Subjects are assigned to each method/tool such that results are unbiased and can be analysed using standard statistical techniques
- as *a case study* where each method/tool under investigation is tried out on a real project using the standard project development procedures of the evaluating organisation
- as *a survey* where staff/organisations that have used specific methods or tools on past projects are asked to provide information about the method or tool. Information from the method/tool users can be analysed using standard statistical techniques.

Although the three ways of organising an evaluation are usually associated with quantitative investigations, they can equally well be applied to qualitative evaluations.

In all, DESMET identified nine distinct evaluation methods including three quantitative evaluation methods, four qualitative evaluation methods and two hybrid methods:

1 Quantitative experiment: an investigation of the quantitative impact of methods/tools organised as a formal experiment.
2 Quantitative case study: an investigation of the quantitative impact of methods/tools organised as a case study.
3 Quantitative survey: an investigation of the quantitative impact of methods/tools organised as a survey.
4 Qualitative screening: a feature-based evaluation done by a single individual (or cohesive group) who not only determines the features to be assessed and their rating scale but also does the assessment. For initial screening, the evaluations are usually based on literature describing the software method/tools rather than actual use of the methods/tools.
5 Qualitative experiment: a feature-based evaluation done by a group of potential users who are expected to try out the methods/tools on typical tasks before making their evaluations.
6 Qualitative case study: a feature-based evaluation

performed by staff who have used the method/tool on a real project.

7 Qualitative survey: a feature-based evaluation done by people who have had experience of using the method/tool, or have studied the method/tool. The difference between a survey and an experiment is that participation in a survey is at the discretion of the subject.
8 Hybrid method 1: Qualitative effects analysis: a subjective assessment of the quantitative effect of methods and tools, based on expert opinion.
9 Hybrid method 2: Benchmarking: a process of running a number of standard tests using alternative tools/methods (usually tools) and assessing the relative performance of the tools against those tests.

Nine methods of evaluation might be regarded as an embarrassment of riches because they lead directly to the problem of deciding which one to use. Thus DESMET also had to identify criteria that would help an evaluator select the most appropriate evaluation method for his or her specific circumstances. DESMET identified eight criteria that might affect evaluation method selection.[4] These allowed the project to identify a set of conditions that favoured each evaluation method (see Table 1).

DESMET produced detailed guidelines for undertaking an evaluation exercise for each of three major evaluation methods: formal experiments, quantitative case studies and feature analysis.

Evaluating DESMET

The DESMET project consortium included two commercial organisations tasked with evaluating the methodology (GEC-Marconi and BNR-Europe). In addition, the University of North London performed evaluations of the guidelines for formal experiments. In terms of project resources, more effort was allocated to evaluating the methodology than was allocated to developing it.

Basis of DESMET evaluation

When we started the DESMET project, we assumed that we could evaluate our methodology simply by using it to conduct some empirical evaluation exercises. However, when we identified the different types of evaluation and their different advantages and limitations, we realised that evaluating DESMET would be more difficult than we had first thought.

Thus, in order to identify an appropriate method of evaluating the DESMET methodology, we attempted to apply our guidelines for evaluation method selection to the DESMET methodology itself. The major problem was to define the nature of the impact of using the methodology. In the view of the DESMET project, the aim of an evaluation methodology is to *reduce the risk* of an invalid or incorrect evaluation exercise. Thus, DESMET fell into the category of methodologies whose impact is not directly measurable on a single project.

Such methodologies cannot be evaluated by quantitative case studies, because a single outcome cannot be used to give a quantitative measure of probability. Surveys can be used to assess methodologies where the measurable impact is indirect, if the methodology is in widespread use. However, DESMET is a new methodology, so a survey was ruled out. Formal experiments were ruled out for three reasons:

- our resources (in terms of effort and timescales) were insufficient to allow for replicating evaluation exercises
- an evaluation methodology has too many elements to fit well into a formal experiment
- it was difficult to identify a control treatment to act as the basis for comparisons

Thus, we used feature analysis to evaluate the DESMET evaluation methodology. As we wanted to demonstrate DESMET in action, we used a case-study-based feature analysis.

Evaluation strategy

DESMET is a multi-component methodology comprising guidelines for performing a number of different evaluation activities (evaluation method selection, quantitative case studies, quantitative experiments and feature analysis). In addition, there are several other supporting components to the methodology (a procedure for assessing evaluation maturity of an organisation; guidelines for coping with human factor issues; a measurement handbook; and a data collection tool). The DESMET partners responsible for evaluating the methodology, therefore, developed a common strategy for evaluating any DESMET component and a standard feature-based questionnaire (called a generic evaluation plan) for each different component.

Generic evaluation plans were prepared for each of the individual components within the methodology. Evaluation criteria were identified for three successive *levels of alidation* defined as follows:

1 *Basic:* Is the component description complete, understandable, usable, internally consistent etc.? Would potential users have confidence that they could use it 'for real' or carry out a trial?
2 *Use:* Is the component 'helpful'? That is, when it was used, did it achieve its objective, produce the specified results, produce usable and relevant results, behave as expected, and not require expert assistance?
3 *Gain:* Is it better than what was available previously (the 'control' situation), i.e. lead to better decision. When there was no control situation against which to test a component, we agreed that the gain validation of the procedure would have to be in terms

of its usability and effects on people's thinking about how to do evaluations.

The generic evaluation plans attempted to provide a context for planning an evaluation in which the methods and procedures described in the component were related to the evaluation criteria. The plans also provided a format for gathering data (via a series of questionnaires) at significant points during an evaluation exercise (called evaluation milestones). Anyone planning an evaluation could use the generic evaluation plan to produce a specific evaluation plan appropriate both to the DESMET component and the host project environment (i.e. the method/tool evaluation that was being performed according to the DESMET evaluation methodology guidelines). This ensured that the evaluation results would be comparable with any other validation of that component.

Having prepared generic evaluation plans for all the components, we identified three evaluation activities:

1 *A review:* This consists of a detailed inspection of, and critical commentary on, the component documentation by an expert reviewer including the completion of an evaluation questionnaire. It is the least satisfactory means of evaluating the benefit of a component and was given least weight in overall component assessment.
2 *Simulated case studies:* This involves an expert reviewer giving consideration to performing a contrived trial by working through the generic validation plan in detail and responding to the questionnaires at the appropriate evaluation milestones. A 'retrospective' validation is a type of simulated case study, where the reviewer answers the questionnaires after a real evaluation has been completed.
3 *Case studies:* This involves applying a DESMET component to a real evaluation exercise, using the generic validation plan to create a feature analysis questionnaire.

The extent of the evaluation activity that took place for each main DESMET component is summarised in Table 2. The Table counts include only reviews and paper trials that were performed by external experts. (Numerous reviews were done within the DESMET project as a part of component development.)

Table 2 Evaluation exercises performed on main DESMET components

Component name	Activity	Occurrence
case study guidelines	case study	5
	simulated case study	1
	review	1
formal experiment guidelines	case study	2
	simulated case study	1
	review	2
feature analysis	case study	1
	simulated case study	1
	review	1
evaluation method selection	case study	4
	limited simulated case study in support of review	4
	review	2

Evaluation results

The most important evaluation exercises were the case studies of the DESMET case study and feature analysis guidelines and these are discussed below. According to the DESMET case study principles, the case studies of the DESMET guidelines were performed on real evaluation exercises. DESMET identified a number of phases that take place in any evaluation project:

1 Context setting: identification of the evaluation goals and constraints.
2 Planning and design: identification of the appropriate evaluation method and planning the evaluation exercise.
3 Preparation: undertaking any pre-evaluation activities (e.g. trials of feature analysis questionnaires, or experimental materials).
4 Execution: performing the evaluation exercise.
5 Data analysis: analysing the results of the evaluation exercise.
6 Dissemination and decision making: using the analysis results to make decisions about technology adoption.

The DESMET evaluation exercises were organised so that an evaluation of the DESMET guidelines was made at the end of each phase of the host evaluation project. Thus all the host evaluation projects contributed to the evaluation of the DESMET methodology even though only four of the six were completed within the DESMET project timescales. In one case, where the intention was to compare two case tools, the evaluation exercises were abandoned after the initial planning stages because of changes in organisational priorities. In another case which was intended to assess the impact of introducing the Schlaer-Mellor method into a department with no previous experience of object-oriented methods, the results of the evaluation were expected some time after the end of the DESMET project.

The other three quantitative case studies and the feature analysis case study are discussed in more detail

Table 3 GUI case study results

Tool	Size (source statements)	Relative code and test effort	% CPU usage
new tool unoptimised	1100	1·0	900 (estimated)
new tool optimised	1600	1·3	90
XLib	800	0·7	47

below. For commercial reasons, details of the methods and tools being investigated are not presented. It must also be noted that case studies are *context dependent evaluation methods*, so there is no implication that another evaluation of the same technology by a different company with different goals would obtain the same results.

DESMET Case Study 1: Assessing a method of reuse by means of a quantitative case study

GEC-Marconi wanted to investigate the effect of a method of reusing specifications. The company followed the case study guidelines for a 'within-project baseline' case study. This case study is used when a method or tool can be applied to a number of functional components within a single project. In this case, product components were monitored during requirements specification and the following attributes were measured:

- percentage of component functionality reused
- component size measured in number of requirements
- number of defects found during a review of requirements
- complexity of component based on a subjective assessment.

Data was collected on 13 different functional areas which were subject to different amounts of reuse. Analysis of the data found no relationship between the amount of reuse and productivity (effort per requirement) or quality (defects per requirement). This suggested that the method of reuse was ineffective. The result of the case study was to abandon this method of specification reuse.

DESMET Case Study 2: Evaluation of alternative user interface tools by means of a quantitative case study

GEC-Marconi had been using X-windows to develop the graphical user interface (GUI) for some software applications. The company wanted to know whether using a more sophisticated tool would be more productive while maintaining the performance characteristics of the GUI. They organised the evaluation as a quantitative case study using a replicated product design. Three versions of a new user interface were produced in parallel: one using the standard X-windows facility, one using the new GUI tool without any attempt to optimise the resulting code, and the third using the new GUI tool to produce optimised code. The results are shown in Table 3. The percentage CPU usage was estimated for the unoptimised code because the user interface could not be made to execute. In this case the standard X-windows facilities outperformed the more sophisticated tool. GEC-Marconi decided not to adopt the new tool.

Table 4 Testing method evaluation results

Method	% defects found	% defects not found	Average module effort
new	60	40	84 to 43
old	50	50	1800

DESMET Case Study 3: Evaluation of automated test generation using a quantitative case study

BNR-Europe was investigating alternative methods of testing. The method currently in use was to generate test cases from English language descriptions of software components. BNR-Europe had developed a new method of generating test cases from a formal description language. Following the DESMET guidelines the company organised its evaluation as a replicated product case study. Five existing software components were seeded with errors and were scheduled for testing using the existing method and using the new method. Testing using the new tool was to be performed by the research group, testing using the existing method by the software developers. During the timeframe of the DESMET project, three modules were tested using the new method and one of the modules had been tested using the existing method. The results are shown in Table 4.

The two effort values for the testing activity identify the impact of re-engineering the formal descriptions in order to generate test cases, compared with the testing effort that would be needed if the formal representation already existed. These results were considered as strong support for the new method, but BNR wants to perform further investigations to ensure that the results generalise beyond the specific modules used in the evaluation exercise.

DESMET Case Study 4: Evaluation of risk management tools by means of feature analysis

GEC-Marconi wanted to identify a risk management tool that could be used both for in-house project management and for inclusion in the portfolio of products that the company itself markets to third parties. Following the feature analysis guidelines, the evaluator identified a hierarchical feature list appropriate for each requirement. The top level features were:

1 supplier maturity
2 tool maturity
3 cost
4 introduction
5 usability
6 compatibility
7 reports
8 constraints

13 tools that performed risk management were identified by a review of the literature, and assessed against the features. The assessment was done twice: once for the

purpose of scoring the tools against the requirements for an in-house tool and once against the requirements for a tool to market to third parties. The feature analysis results allowed the selection of one tool that scored best for both purposes.

Evaluation conclusions

The most important aspects of the DESMET evaluation were case studies of the different components of the methodology on real evaluation exercises, although we also initiated reviews and simulated case studies. The case studies covered: specification methods, GUI evaluation, reuse, risk tools and testing methods. This wide range of methods/tools gave us some confidence that we had succeeded in our aim to develop a general evaluation method.

In general, the results of the evaluation exercises were favourable.[5] In particular, the GEC-Marconi evaluators found that the methodology provided guidance, enabled hypothesis testing, and led to clear results which were used to support decision-making. However, we are not complacent about the results of the evaluation exercises since there were some fairly serious criticisms of the guidelines we produced, mainly concerned with the presentation of information in the original research reports. As a result we have been upgrading and improving the methodology since the end of the project. Perhaps the results of the evaluation exercise are best summed up in the words of Niall Ross of BNR.

'DESMET will give you useful information, show you some interesting examples and ask some interesting questions you might not have thought of yourself, but whether they stimulate you to produce valuable questions depends on your native wit.'

Adoption of DESMET

Adoption of new methods is slow, but so far adoption of the DESMET evaluation methodology has been encouraging. After a public workshop in June 1994 at the end of the project, components of the methodology in the format of technical reports from the DESMET project were purchased by 16 organisations, as well as being made available on a restricted basis to various university researchers. A follow-up survey in June/July 1995 managed to contact 12 organisations and of those eight were using the results. The users came from a variety of industrial sectors and vary from small companies to very large organisations. All reported that they found the methodology to be both usable and valuable.

In addition, the two commercial partners in the DESMET project (BNR-Europe and GEC-Marconi) are using the technology. The University of North London has included the DESMET methodology in its teaching and research programmes. Also, Southampton University which assisted in the evaluation of the DESMET project continues to use the results.

Conclusions

DESMET is a comprehensive methodology for assisting the organisation of software method/tool evaluation exercises. The DESMET methodology has itself been subjected to an extensive evaluation exercise. The evaluation of the DESMET methodology by the methodology itself provides a unique insight into the evaluation of methods. As a result of our experiences, we are strongly of the opinion that all developers of new methods should undertake evaluation exercises *as a natural part of method development.*

Practical use has been made of the DESMET methodology by a number of organisations; some outline examples which illustrate this are included in this article. It provides a sound basis for the evaluation of tools and methods enabling informed decision making to take place.

Information about the DESMET methodology can be obtained from a number of publicly available articles. These are updated versions of the original research reports that take into account many of the problems identified by the evaluation exercise. Shari Lawrence Pfleeger reported the guidelines for quantitative experiments in References 6 and 3; Kitchenham *et al.* have reported the guidelines for quantitative case studies[7] and there is a series of articles currently appearing in *SIGSOFT Notes* that discuss evaluation method selection, managerial and social issues affecting evaluation exercises, feature analysis, case studies and the evaluation of DESMET.[2] The metrics handbook produced by the DESMET project has formed the basis of a recent book.[8]

References

1 FENTON, N., LITTLEWOOD, B., and PAGE, S.: 'Evaluating software engineering standards and methods', *in* THAYER, R., and McGETTERICK, A. D. (Eds.): 'Software engineering: a European perspective' (IEEE Computer Society Press, 1993)

2 KITCHENHAM, B. A.: 'Evaluating software engineering methods and tools. Parts 1 to 12', *SIGSOFT Notes*, starting January 1996

3 PFLEEGER, S. L.: 'Experimental design and analysis in software engineering. Parts 1 to 5', *SIGSOFT Notes*, 1994 and 1995

4 KITCHENHAM, B. A., LINKMAN, S. G., and LAW, D. T.: 'Critical review of quantitative assessment', *Software Engineering Journal*, 1994, **9**, (2), pp.43-53

5 SADLER, C., PFLEEGER, S. L., and KITCHENHAM, B. A.: 'Trials results and validation', DESMET Report D4.1, September 1994 (currently not in the public domain)

6 PFLEEGER, S. L.: 'Experimental design and analysis in software engineering', *Annals of Software Engineering*, 1995, **1**, (1)

7 KITCHENHAM, B. A., PICKARD, L. M., and PFLEEGER, S. L.: 'Case studies for method and tool evaluation', *IEEE Software*, July 1995, **12**, (4), pp.52-62

8 KITCHENHAM, B. A.: 'Software metrics: measurement for software process improvement' (NCC Blackwell Publishers, 1996)

Barbara Kitchenham and Steve Linkman are now members of the Computer Science Department, Keele University. During the DESMET project, they worked for NCC as senior researchers on the project. David Law is an independent consultant. He was the project manager of the DESMET project.