Costs and Benefits of Software Process Improvement[1]

Khaled El Emam and Lionel Briand
Fraunhofer - Institute for Experimental Software Engineering

Introduction

In recent years a substantial number of organizations have gained experience in software process improvement (SPI). Furthermore, some researchers have studied such organizations by collecting and analyzing costs and benefits data on their SPI efforts. The objective of this report is to review and summarize the empirical evidence thus far on the costs and benefits of SPI. The intention is that this review would be utilized to support the business case for initiating and continuing SPI programs, to aid in the selection amongst the alternative improvement paradigms, to make more accurate estimates of the costs and benefits of such efforts, and to help set and manage the expectations of technical staff and management.

The need for such a review is supported by the results of two recent surveys that were conducted by the SEI. The first survey was administered to individuals at the National SEPG Conference in 1993 and at an SPI tutorial during the Software Engineering Symposium in 1993 [25]. The respondents represented organizations that had mature SPI programs. More than seventy percent stated that they need information on the benefits of SPI (by choosing the "very high" or "high" response category in terms of characterizing their needs), which was also ranked as the highest need of the respondents. This indicates a need for consolidation of the evidence on the benefits of SPI. The second survey solicited information from organizations that had conducted software process assessments between 1992 and 1993 [26]. The results indicate that 77% of the respondents "Strongly Agree" or "Agree" that SPI has taken longer than expected and 68% stated that SPI has cost more than expected. This indicates a need for information to help estimate the costs of SPI and to set and manage expectations from SPI.

Two general paradigms to SPI have emerged, as described by Card [10]. The first is the analytic paradigm. This is characterized as relying on *"quantitative evidence to determine where improvements are needed and whether an improvement initiative has been successful"*. The second, the benchmarking paradigm, *"depends on*

---

[1] This appears as International Software Engineering Research Network technical report ISERN-97-12, 1997.

identifying an 'excellent' organization in a field and documenting its practices and tools". Benchmarking assumes that if a less-proficient organization adopts the practices of the excellent organization, it will also become excellent. These SPI paradigms are covered in this report. Readers can select the data that is most applicable to the SPI paradigm that they intend to use or are using.

Empirical studies that have been conducted do not answer all of the questions about SPI; those that have been answered not to the level of detail that some may wish. However, the available data do provide us with credible guidance in our SPI efforts, which is undoubtedly preferable to no guidance.

As the title suggests, the report is divided into two main parts. The first presents data on the costs of SPI, and the second on the benefits of SPI. In the second part we also use results from the empirical research literature to provide some guidelines to help attain the promised benefits of SPI. Subsequently, we discuss some methodological issues pertinent to benefits studies in order to give the reader an appreciation of the issues involved in doing this kind of work and also to help interpret future benefits data.

The Costs of Software Process Improvement

The most common and interpretable measures of the costs of SPI are in terms of dollars and/or effort. A recent study sponsored by the US Air Force [6] found that government organizations tend to characterize investments in process improvement in terms of costs, whereas industry tend to characterize it in terms of effort expended on SPI activities. In some cases, cost measures such as calendar months have also been used. The studies that we summarize below show the costs of SPI using different analytical and benchmarking approaches. The amount of detail that we can present is directly a function of the amount of publicly available information.

Ref.	Organization & SPI Program	Costs	Benefits
[30]	• SPI effort at the Software Engineering Devision of Hughes Aircraft • The division had 500 professional employees at the time	• The assessment itself cost US$45,000 • Cost of a 2 year SPI program was US$400,000 • Implementation of the action plan to move from ML1 to ML2 was 18 calendar months	• Achieved annual savings of US$2M • Benefits were calculated to be 5 times the improvement expenditures • Quality of work life had improved (e.g., fewer overtime hours by the software engineers)
[44]	• SPI effort led by the Schlumberger Laboratory for Computer Science	Large engineering centers (120-180 engineers) have 1-5 full-time staff on SPI Smaller centers (50-120 engineers) have up to 3 full-time staff on SPI	• Improved project communication • Customer reports confirm that product quality has improved • One group improved time to market by reducing requirements validation cycles to 15 from 34 • One group more than doubled its productivity • One group increased the percentage of projects completed on schedule from 51% to 94% • One group almost halved the defect density in its products
[6] [7]	• Data was collected from 33 companies using questionanires and/or interviews.	• The authors present examples of data on the costs of activities related to SPI. • For example, some organizations increased from 7% to 8% of total effort on data collection, and increase upto 2% of project costs on fixing design defects.	• Some organizations witnessed increased productivity, reduced defect levels, reduced rework effort, reduction in costs and greater within estimate project completions. • In particular, some organizations achieved a ROI of 10:1 • Other benefits included less overtime and employee turnover, and increased cooperation between functional groups
[8]	• Corporate-wide SPI effort at AlliedSignal Aerospace starting in 1992	Using data on SEPG investment measured in full-time equivalent headcount for 8 sites, the maximum was 4%	• One site realized a 7:1 productivity increase in the calendar time to generate documents over 1000 pages with a 50% reduction in the cost per page • Independent V&V deficiency reports on the documentation have decreased to approximately zero • One site reported that the LOC maintained per person has increased by 50% and testing time has decreased by 60% with no evident increase in delivered defects
[14]	• Organization is the Software Systems Laboratory in Raytheon, employing 400 software engineers • SPI initiative started in 1988; results reported after five years • Organization has progressed from Level 1 to Level 3 during that period	US$1 million invested every year	• A 7.7:1 return on every dollar invested • Rework costs reduced from 41% of overall project costs to 11% • More projects finish ahead of or on schedule and under or on budget • Productivity increases by a factor of 2.3 in 4.5 years • Software engineers spend fewer late nights and weekends on the job and improved general morale

Figure 1: Organizational experiences illustrating the costs and benefits of SPI.

Costs of Assessment and Improvement Based on the CMM

A number of companies have publicized the cost details of their process improvement efforts based on the CMM. Some of these are summarized in Figure 1. Another study conducted at the SEI determined the amount of time it takes organizations to increase their maturity levels on the CMM for the first three levels [28]. The distribution of assessments that used the original SPA method and the replacement CBA-IPI method in the data set is not clear however, and whether any differences in method would have had any effect on the time it takes to move up one maturity level.

Two groups of organizations were identified: those that moved from level 1 to level 2, and those that moved from level 2 to level 3. On average, it takes organizations 30 months to move from level 1 to level 2. Those organizations, however, varied quite dramatically in the amount of time it takes to move up one maturity level. A more outlier resistant measure would be the median. In this case, the median was 25 months. Organizations that moved from level 2 to level 3 took on average 25 months (the median was also 25 months).

It is expected that the size of the organization would have a significant impact on the number of months it takes to move from one maturity level to another. The variation in the size of the organizational units that were assessed was not given in the report however. Therefore, these results should be taken as general guidelines to check an organization's own estimates of the time it takes to move up the maturity ladder.

Another study of US companies found results that are consistent with those mentioned above [6]. It was found that organizations at level 2 spend between 12 to 36 months at level 1with an average of 21 months, and organizations at level 3 had spent 22-24 months at level 1 with an average of 23 months. Organizations at level 3 spent from 12 to 20 months at level 2 with an average of 17.5 months. This is corroborated with data from the improvement efforts at AlliedSignal [8] where advancement from Level 1 to 2 and from Level 2 to Level 3 took 12-14 months across different sites.

Costs of Assessments Based on SPICE

During the SPICE trials, data was collected on amount of effort it takes to conduct an assessment. The median value for an assessment was found to be 110 man hours. This was based on data collected from the ratings of 324 process instances during the SPICE trials. Variation, however, was quite substantial ranging from approximately 33 person-hours to 824 person-hours. These numbers include total assessor and assessee effort. Of course, one explanation for this is the fact that some assessments rated much more process instances than others.

The box-plots for the assessor and assessee effort per process instance rated are shown in Figure 2. The median total effort spent by the assessment team members per process instance is 8.3 person hours. This number excludes the effort to fill up the forms and questionnaire that were required by participants in the SPICE trials. However, the overall variation is very large. The minimum value is 1.4 person-hours per process instance and the maximum was 100 person-hours. The median total effort spent by the employees of the assessed organizational unit per process instance was 7.6 person-hours. The variation again was large, ranging from a minimum of 0.7 person-hours per process instance to a maximum of 40 person hours.

On-going work by the SPICE trials team focuses on explaining the variation that was seen in this data. However, the current results can still be utilized as an aid in estimating assessment effort.

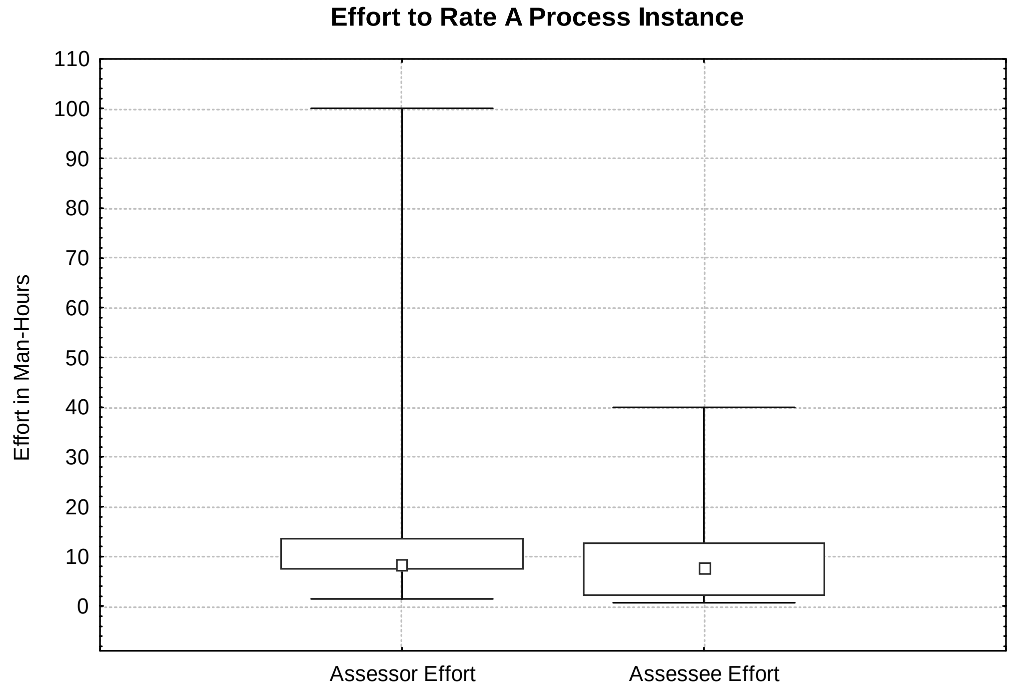

Figure 2: Effort in person-hours to judge and rate the adequacy of a process instance for each of the five SPICE process categories.

Costs of Registration to ISO 9001

A multiple regression model has recently been constructed to estimate the effort it would take an organization to meet the requirements of ISO 9001 [40]. Data was collected from 28 software organizations that were registered to ISO 9001 in Canada and the USA. There are two inputs to the model: (a) the size of the organization in number of employees, and (b) the degree of non-compliance to ISO 9001 clauses. Both sets of data were collected by questionnaire and a sample of responses were verified with the respondents to increase confidence in the reliability of the responses. The model to predict effort in man-months is:

$$Ln\ (effort) = -2.793 + 0.692 * Ln\ (x_1) + 0.74 * Ln\ (x_2)$$

where:
x_1 = number of employees within the scope of registration
x_2 = degree of compliance of the organization to the ISO 9001 clauses prior to the improvement effort

The model was validated using data collected from five organizations that were not included in the model development sample. A brief comparison of the model prediction versus the actual effort is given inFigure 3.

Org. #	Size	Non-compliance (%)	Predicted	Actual	Residual
1	175	35%	30.3	31.2	0.9
2	108	15%	11.6	13	1.4

3	170	30%	26.5	27	0.5
4	45	100%	25.8	36	10.2
5	100	70%	34.4	37	2.6

Figure 3: Comparison of actual versus predicted effort for ISO 9001 registration.

Costs of Measurement-Based Process Improvement

Measurement-based process improvement relies on the acquisition of data and factual information in order to suggest process improvement actions. As opposed to the models presented above, it does not rely on a generic model of software development and pre-defined improvement steps (an overview of this approach to SPI is given in [4]). In the context of measurement-based process improvement, another advantage is that precise cost information is usually readily available and can therefore be used for computing quantitative and objective cost estimates. However, it is difficult to determine to which extent the cost figures below are comparable to the ones presented for CMM, SPICE, and ISO-9001. The cost of measurement implies a continuous support to projects and not just a one-time snapshot of the software processes in an organization. In addition, the benefits of measurement-based process improvement also include a better manageability of projects as well as supporting process improvement activities. It also provides a much more detailed characterization of processes and, more importantly, provide insights into the products of these processes.

In order to assess the cost of measurement-based process improvement, we will use four main sources for which data are in the public domain: NASA GSFC Software Engineering Laboratory (SEL) [2], Motorola [12], Hewlett-Packard [27], and Philips Sound & Vision [41]. The results seem to be consistent across these four organizations. They all seem to indicate that the cost of measurement itself (collecting, checking, storing, and analyzing data) should not exceed 1 or 2 percent of the project resources. However, additional effort has to be planned for technology transfer if an organization wants to make improvement possible. A careful introduction and tailoring of new technologies will take substantial effort but we do not believe there is any other alternative for improvement.

At the NASA GSFC SEL, about 10% of the cost of development is invested in SPI activities including research (designing studies and analyzing results), technology transfer (producing standards and policies), and data processing (collecting forms and maintaining databases), the latter accounting for less than 2% of the cost. At Motorola, data collection activities represent roughly 1% of project resources. Philips Sound & Vision reports a cost of approximately 1%, including both data collection and the introduction of inspections. The discrepancy between these figures is in part due to the strong involvement of the NASA GSFC SEL in research activities related to measurement-based SPI and to the fact that these four organizations did not report cost in a consistent manner.

Ref.	Description of Study	Results
[26]	• Survey of individuals whose organizations have been assessed against the CMM • The authors tested the relationship between maturity levels and subjective measures of effectiveness.	For the first three maturity levels, organizations at higher maturity levels tend to perform better on the following dimensions (respondents chose either the "excellent" or "good" response categories when asked to characterize their organization's performance on these dimensions): • ability to meet schedule, • product quality • staff productivity • customer satisfaction, and • staff morale. The relationship with the ability to meet budget commitments was not found to be statistically significant.
[35]	• correlational study that investigated the benefits of moving up the maturity levels of the CMM • They obtained data from historic U.S. Air Force contracts. Two measures were considered: (a) cost performance index which evaluates deviations in actual vs. planned project cost, and (b) schedule performance index which evaluates the extent to which schedule has been over/under-run.	• generally, higher maturity projects approach on-target cost • generally, higher maturity projects approach on target schedule

Figure 4: Summaries of the benefits of higher maturity level scores on the CMM.

The Benefits of Software Process Improvement

The types of studies that document the benefits of SPI include organizations that have used the CMM, organizations that have used ISO 9001, and organizations that have followed other models. Benefits data is also available for organizations that have followed measurement based improvement programs.

Benefits of Improvement Based on the CMM

Examples of organizations that have documented the benefits of improvement based on the CMM are summarized in Figure 1: Organizational experiences illustrating the costs and benefits of SPI.
. Studies that have investigated the benefits of higher maturity level scales on the CMM based on a statistical analysis of data from a larger number of organizations are summarized in Figure 4.

Benefits of Registration to ISO 9001

Many software organizations are being audited against the clauses of ISO 9001. A number of surveys have been conducted that evaluate the benefits of ISO 9001 registration in industry. Some of the results of these surveys have been presented in [43] and are summarized in Figure 5.

With respect to registration to ISO 9001, the few studies that have been conducted seem consistent in their findings that it can bring benefits. However, many of these studies were not specific to software organizations. Therefore, more research specifically with software organizations would help the community better understand the effects of registration.

Description of Survey	Overview of Some Relevant Findings
One survey conducted in 1993 had 292 responses with almost 80% of the responding organizations being registered to ISO 9001.	• 74% felt that the benefits of registration outweighed the costs • 54% received favourable feedback from their customers after registration
A survey of companies in the U.K. had 340 responses from companies that were registered.	• It was found that 75% of the respondents felt that registration to ISO 9001 improved their product and/or service.
A survey of companies that were registered in the U.S.A. and Canada with 620 responses.	• The most important internal benefits to the organization included: better documentation (32.4%), greater quality awareness (25.6%), a positive cultural change (15%), and increased operational efficiency/productivity (9%); and • The most important external benefits to the organization included: higher perceived quallity (33.5%), improved customer satisfaction (26.6%), gaining a competitive edge (21.5%), and reduced customer quality audits (8.5%).
A survey of 45 software organizations in Europe and North America that have become ISO 9001 registered	• 26% reported maximum benefit from increased efficiency • 23% reported maximum benefit from increased product reliability • 22% reported maximum benefit from improved marketing activity • 14% reported maximum benefit from cost savings, and • 6% reported maximum benefit from increases exports

Figure 5: Surveys of the benefits of registration to ISO 9001 (source [43]).

Benefits of Improvement Using Other Models

Two studies have evaluated models that measure the maturity of software organizations. The two studies that we review are summarized in Figure 6.

Ref.	Description of Study	Overview of Relevant Findings
[17]	A questionnaire was developed to measure the maturity of Management Information Systems (MIS) organizations along four orthogonal dimensions. These four dimensions were: (a) standardization, (b) project management, (c) tools, and (d) organization. The authors investigate the relationship between maturity and the success of the requirements engineering process (RE success). Two dimensions of RE success were measured [18]: the quality of RE service, and the quality of RE products.	The relationship between the organization dimension and the quality of service was found to be moderate (a Pearson correlation coefficient of 0.58) and was statistically significant. The relationships with the quality of RE products were all small or non-existent and not statistically significant.
[32]	Jones presents the results of an analysis on the benefits of moving up the 7-level maturity scale of Software Productivity Research (SPR) Inc.'s proprietary model. This data were collected from SPR's clients.	His results indicate that as organizations move from Level 0 to Level 6 on the model they witness (compound totals): • 350% increase in productivity • 90% reduction in defects • 70% reduction in schedules

Figure 6: Studies investigating the benefits of SPI using models and methods other than the CMM and ISO 9001.

Benefits of Measurement-Based Improvement

With respect to measurement-based software process improvements, the benefits reported in the literature vary significantly. This is to be expected because the impact of measurement-based process improvement will depend on the inital level of maturity of the organization, the complexity of the systems under development, and varying factors over the period of measurement. The results seems to suggest very significant improvements both in terms of productivty and quality. A summary of the benefits is given in Figure 7. Such levels of improvement obviously outweigh costs such as those described earlier in this report and therefore demonstrate the cost effectiveness of measurement-based process improvement. But measurement, like assessment, does not create improvement. It just makes it possible and supports it. For example, investing in measurement and neglecting technology transfer would not be likely to pay off as a process-improvement strategy. This is why these results should be interpreted with care since they do not imply that measurement in itself will lead to improvement.

Description of Study	Overview of Relevant Findings
SPI efforts at NASA GSFC Scope: flight dynamics division (FDD) Data that has been collected over a period of 20 years and 100 projects at the NASA GSFC SEL. The complexity of the flight dynamics software developed has tremendously increased over the years. Around 300 software engineers in the FDD Use the G/Q/M approach to measurement	• Cost to deliver decreased 58% in 5 years due mainly to an increase in reuse • Error rates per KSLOC decreased by 35% in ten years (from 8.4 to 5.3 errors/KSLOC) • Improved ability to predict, control and manage the cost and quality of software being produced.
SPI efforts at Hewlett-Packard Scope: Company-wide initiative	• 3 fold increase in productivity over 4 years • 80% reduction in defect density over 4 years • reduction of the number of major defects during postrelease • predict testing completion within 10% of its actual duration and effort
SPI effort at Philips Sound & Vision Scope: data collected on three projects representing an effort of 60 staff-years Software for consumer electronics products CMM level 2 organization	• saved 10 percent of effort through early detection of defects • reduced life cycle time
SPI effort at Motorola Scope: Company-wide initiative 1 division of 350 software engineers 1 division of 70 engineers Use the Q/G/M approach to measurement	• 1 division achieved a 50 times reduction in defect density over 3.5 years. • Significant cost reduction due to improved quality • Better project management, e.g., Improved ship-acceptance criteria and schedule estimation accuracy • ÑIn general, the overall cost is acceptable and justified.ì

Figure 7: Benefits of measurement-based improvement.

Attaining the Benefits from Process Improvement

Strong business interests by suppliers and users of software process improvement models and methods demand that there is empirical evidence demonstrating benefits. The above review of the empirical evidence supports the contention that software process does matter. The individual organizational experiences show that SPI can increase effectiveness, and the surveys show that on average organizations that do implement what are believed to be good software processes are better than those that do not or do less. However, attaining the benefits of SPI is not necessarily a simple matter of implementing a list of processes or process management practices. A detailed analysis of the empirical literature helps provide some guidelines to consider while pursuing an SPI effort.

Evidence for Customizing Improvement Efforts

The surveys reviewed earlier seem to show that substantial benefits would be gained from SPI. But, do organizations that focus on process always benefit? Existing evidence suggests that the extent to which an organization’s effectiveness improves due to the implementation of good software processes or software management practices is dependent on the characteristics of the project(s) and the organization. However, the overall evidence remains equivocal as to which factors moderate the relationship between process and effectiveness.

For example, in [17] the relationship between some dimensions of maturity and the success of the requirements engineering process was investigated. As summarized in Figure 6, it was found that only one dimension of maturity was related to success. This may indicate that the relationship is moderated (e.g., the magnitude of the relationship is different for large vs. small organizations). A number of possible moderating variables were considered.

One possible moderating variable is the size of the MIS organization. For example, there have been some concerns that the implementation of some of the practices in the CMM, such as a separate Quality Assurance function and formal documentation of policies and procedures, would be too costly for small organizations [5]. Therefore, the implementation of certain processes or process management practices may not be as cost-effective for small organizations as for large ones. To investigate the possibility that benefits depend on the size of the organization, the sample of MIS organizations was divided into those that were small (less than 100 employees) and those that were large (100 or more employees). Then the correlations between maturity and RE success were compared for the small and large MIS organizations. This analysis shows that there are no differences in the correlations between small and large organizations for all the dimensions of maturity. Therefore, MIS organization size does not seem to moderate the relationship. This result is consistent with that found in [26] for organization size and [13] for project size, but is at odds with the findings from [5].

To further confuse the issue, an earlier investigation [37] studied the relationship between the extent to which software development processes are standardized and MIS success. It was found that standardization of life cycle processes was associated with MIS success in smaller organizations but not in large ones. This is in contrast to the findings cited above. Therefore, it is not clear how organization size moderates the benefits of process and the implementation of process management practices.

Another possible moderating variable is the business sector of the organization. One study on the benefits of higher CMM maturity did not find differences in terms of benefits for different industrial sectors [26]. Another study that investigated the effects of process implementation on meeting schedule and budget targets and on product quality did not, in general, find different effects for military vs. non-military projects [13]. An alternative differentiation is between government and non-government organizations. Using the data set in [17], the organizations were divided, but this time depending on whether they were government or not. Then the correlations between maturity and RE success were compared for these two groups. This analysis indicates that there are no large differences in the correlations between government and non-government organizations for all the dimensions of maturity. Although, on the Project Management dimension, the difference does approach statistical significance (two tailed $p=0.07$), indicating that potentially the relationship between the Project Management maturity dimension and the quality of RE service is larger for government organizations. This indicates that business sector *may have* a small moderating effect on the maturity $\leftrightarrow$ RE success relationship.

Size and industrial sector are not the only factors that may have an effect on the benefits of process improvement. One study investigated the effects of user participation in the requirements engineering process and requirements engineering success [20]. It was found that the extent of uncertainty about information requirements had an impact on the degree to which user participation in the requirements engineering process was beneficial. Therefore, project uncertainty is another factor to consider when determining the benefits of implementing "good" practices.

The implications of the results presented above is that following stipulations about implementing certain processes or process management practices across the board (i.e., irrespective of the organizational and project characteristics) is ill advised. This is so until more consistent empirical evidence can be furnished. Meanwhile, one should evaluate the specific contexts of the organization and projects before selecting process improvement actions and customize their SPI efforts to their local conditions.

Evidence for Considering Non-Process Factors

None of the studies reviewed establishes a causal relationship, i.e., that process improvement is the cause of benefits that are witnessed. To establish causation one must at least rule out other possible causal factors that could have led to the benefits witnessed over the same period. Also, experience reports documenting benefits of SPI would have to rule out natural progress (i.e., if the organization did not make any changes, would they have achieved the same benefits?).

It is clear that implementation of processes or process management practices are not the only factors that will influence effectiveness. Bach [1] has made the argument that individual software engineer capabilities is a critical factor having an impact on project and organizational effectiveness. He even goes further, stating *"that the only basis for success of any kind is the 'heroic efforts of a dedicated team'."* The importance of individual capability is supported by empirical research. For instance, one study found that the capabilities of the lead architect were related to the quality of requirements engineering products [16]. Another study found a relationship between the capability of users participating in the requirements engineering process and its success [21]. Other field studies of requirements and design processes also emphasized the importance of individual capabilities [11][19].

The implementation of automated tools has been advocated as a factor that has an impact on effectiveness. This assertion is supported by empirical research. For instance, one study of the implementation of an Information Engineering toolset achieved increases in productivity and decreases in post-release failures [24].

The best that can be attained with studies that focus only on process factors is strong evidence that SPI is *associated* with some benefits or that organizations *could* benefit from SPI activities. In order to improve our understanding of the influences of other factors on effectiveness more sophisticated empirical studies would have to be conducted. These would include building multivariate models that take the influence of non-process factors into account and investigate the interactions between process and non-process factors. Thus far, most studies have been limited to primarily bivariate analyses.

The message from current research, however, is that there are other non-process factors that do have an impact on organizational and project effectiveness. It would not be prudent to focus only on process and forget everything else. Software Process Improvement should be part of an overall strategy that addresses, at least, weaknesses in people capabilities and the needs for tool support. Furthermore, the success of SPI is strongly influenced by the approach used for implementation of new practices. For example, a good practice may not provide anticipated benefits because the implementation was not performed properly.

Evidence for Goal Directed Improvement

It has been shown that different processes have different impacts on the same measures of effectiveness. For example, one study that examined the effect of four dimensions of organizational maturity on the success of requirements engineering processes found that some maturity dimensions are related to success, while others were not [17]. In particular, the dimensions measuring standardization, project management, and tools were found to be unrelated to success; but the organization dimensions was related to success. Another study [13] investigated the relationship between seven software processes and measures of project performance. The results indicated that some processes and practices, such as project planning and cross functional teams, were related to product quality. However, practices such as user contact and prototyping were not related to quality.

Furthermore, it is not uncommon for studies that investigate the benefits of the implementation of processes to obtain different results depending on the measures of effectiveness that are used. For example, one study found that the use of cross functional teams was related to the quality of products but not to meeting schedule and budget targets [13]. Another study [17] found a relationship between the organization dimension of maturity and the quality of requirements engineering service, but not with the quality of requirements engineering products.

The message from this research is that an organization should identify its business goals, identify measures for evaluating the attainment of its business goals, and then select to implement processes and process management practices that are most likely to have an impact on these measures that are important for the organization. Off-the-shelf generic lists of processes to implement may not be as effective for all organizations and may not address the goals of all organizations.

There are organizations who are taking this assertion seriously in deciding on the factors to focus their improvement efforts on. In an analysis of assessment data from 59 sites representing different business sectors (e.g., DoD contractor and commercial organizations) and different project sizes (from less than 9 peak staff to more than 100) available at the SEI [34], more than half of the sites reported findings that do not map into the KPA's of the CMM. This indicates that organizations are identifying issues to be addressed not covered by the CMM. Another implication of the above assertion is the right order which organizations should improve their processes. In a report of SPI based on the CMM [8] it is noted that *"Business demands often necessitate improvements in an order which is counter to the CMM."* In that particular case, the organization initiated some process improvements that were not necessarily congruent with their next level of maturity, but driven by their business objectives. Further evidence against following generic improvement paths comes from a study reported in [15]. The authors investigated whether the maturity path suggested by the process maturity framework of Humphrey and Sweet [31] follows a natural evolutionary progression. Their analysis was based on the basic idea that questions representing maturity levels already passed by organizations would be endorsed (i.e., scored yes) while items representing maturity levels not reached would fail. Their results did not support the original maturity path and led the authors to suggest that the original model seemed "arbitrary" in its ordering of practices and is "unsupported". The first five levels o the alternative maturity model that they empirically derived is shown in Figure 8. Of course, further studies are necessary to confirm this alternative model, but at least it enjoys some empirical support thus far. In addition, this study highlights that we, as a community, still do not know the „right“ ordering of practices, and hence the importance of driving SPI along a path based on the organization's objectives, and not necessarily by that of a generic model.

Level 1: Reviews and Change Control

- Is a mechanism used for controlling changes to the code? (Who can make changes and under which circumstances?) (L2)
- Are internal software design reviews conducted? (L3)
- Are software code reviews conducted? (L3)
- Is a mechanism used for controlling changes to the software requirements? (L2)

Level 2: Standard Process and Project Management

- Is a mechanism used for controlling changes to the software design? (L3)
- Does the software organization use a standardized and documented software development process on each project? (L3)
- Do software development first line managers sign off on their schedules and cost estimates? (L2)
- Is a formal procedure used in the management review of each software development prior to making contractual commitments? (L2)
- Is a formal procedure used to produce software development schedules? (L2)
- Are formal procedures applied to estimating software development cost? (L2)
- Is a mechanism used for managing and supporting the introduction of new technologies? (L4)

Level 3: Review Management and Configuration Control

- Are the action items resulting from code reviews tracked to closure? (L3)
- Are the actions items resulting from design reviews tracked to closure?(L3)
- Are the review data gathered during design reviews analyzed? (L4)
- Is there a software configuration control function for each project that involves software development? (L2)
- Are code review standards applied? (L4)
- Is a formal procedure used to make estimates of software size? (L2)
- Is a mechanism used for periodically assessing the software engineering process and implementing indicated improvements? (L4)

Level 4: Software Process Improvement

- Are analyses of errors conducted to determine their process related causes? (L4)
- Is a mechanism used for ensuring compliance to software engineering standards? (L3)

Level 5: Management of Review and Test Coverage

- Are design and code review coverages measured and recorded? (L4)
- Is test coverage measured and recorded for each phase of functional testing? (L4)

Figure 8: Empirically derived maturity model (first 5 levels only).

Threats to the Validity of Benefits Studies

As already shown, there has been substantial empirical research to evaluate the benefits of software process improvement. Undoubtedly, there will be more such empirical research in the future. However, as well as demanding empirical evidence, consumers of empirical research must evaluate these works critically in order to determine how much confidence one should have in their results.

In this section we review a number of methodological issues that threaten the validity of conclusions that can be drawn from empirical studies of software process improvement. This review is intended to achieve two purposes. First, to present guidelines for the consumers of such works to help them evaluate the works. While some of the issues brought up may seem minor to non-specialists, they do in fact have substantial impacts on the conclusions that one can safely draw from the studies. Second, to provide researchers in this area with a list of issues to note while conducting their research, and perhaps contribute to its improvement.

Biases of Particular Evaluation Methods

Two classes of empirical studies of the benefits of SPI have been conducted and reported: *case* studies and *correlational* studies. Case studies describe the experiences of a single organization (or a small number of selected organizations) and the benefits it gained from increasing its maturity level. Examples of case studies are given in Figure 1 as well as in [3][9][29][36][38]. Case studies are most useful for showing the *potential benefits* from the implementation of good processes. Given the substantial number of case studies documenting benefits from SPI, it is clear that it is possible to obtain considerable benefits from SPI. However, SPI case studies have a methodological disadvantage that makes it difficult to generalize their results. Case studies tend to suffer from a selection bias because organizations that have not shown any process improvement or have even regressed will be highly unlikely to publicize their results, so published case studies tend to show mainly success stories (e.g., *all* the references to case studies in this report are success stories). More worrisome is that we do not have an evaluation of how many case studies that are not success stories actually exist but were never published. Therefore, case studies do not demonstrate a general association between SPI and some benefits.

With correlational studies, one collects data from a number of organizations and investigates relationships between the implementation of good processes (e.g., maturity) and organizational and/or project effectiveness statistically. In correlational studies, data is usually collected through sample surveys, although this is not *always* the case. Correlational studies are useful for showing whether a *general association* exists between increased process implementation and effectiveness, and under what conditions. Examples ofcorrelational studies are given inFigure 4 and Figure 6.

One problem is that the majority of organizations do not collect objective process and product data (e.g., on defect levels, or even keep accurate effort records). Organizations following the benchmarking paradigm do not necessarily have measurement programs in place to provide the necessary data. Primarily organizations that have made improvements and reached a reasonable level of maturity will have the actual objective data to demonstrate improvements (in productivity, quality, or return on investment). This assertion is supported by the results in [6] where, in general, it was found that organizations at lower CMM maturity levels are less likely to collect quality data (such as the number of development defects). Also, the same authors found that organizations tend to collect more data as their CMM maturity levels rise. Conversely, organizations following the analytic paradigm will tend to start measurement programs early in their SPI efforts, and therefore potentially have costs and benefits data. However, it was reported in another survey [42] that for 300 measurement programs started since 1980, less than 75 were considered successful in 1990, indicating a high mortality rate for measurement programs. This high mortality rate indicates that it may be difficult right now to find many organizations that have implemented measurement programs.

Therefore organizations that fail in their SPI efforts or who do not progress are less likely to be considered as viable case studies due to the lack of sufficient data. This enforces the case study selection bias alluded to earlier. Also, projects that have low implementation of processes or that do not have successful measurement programs may have to be excluded from a correlational study for the same reason. This would reduce the variation in the variables being measured, and thus reduce (artificially) the coefficients obtained from thecorrelational study.

This particular problem has been addressed in various ways in correlational studies however, but remains an issue for case studies. The study in [35] used data from contracts with the US Air Force where schedule and budget data is regularly collected irrespective of the organization's maturity. The study by Jones relies on the reconstruction of, at least, effort data from memory, as noted in [33]: *"The SPR approach is to ask the project team to reconstruct the missing elements from memory."* The rationale for that is stated as *"the alternative is to have null data for many important topics, and that would be far worse."* The general approach is to show staff a set of standard activities, and then ask them questions such as which ones they used and whether they put in any unpaid overtime during the performance of these activities. For defect levels, the general approach is to do a matching between companies that do not measure their defects with similar companies that do measure, and then extrapolate for those that don't measure. It should be noted that SPR does have a large data base of project and organizational data, which makes this kind of matching defensible. Other studies, such as [17][26] used subjective measures collected via questionnaires, therefore circumventing the difficulties of the collection of objective organizational and project data.

Appropriate Measurement

The manner in which variables are measured can have a non-trivial impact on the results of a study. Ideally, depending on the type of measure, appropriate measurement procedures should be followed. Below we discuss two common measurement problems in benefits evaluation studies.

Studies that utilize measures involving subjectivity should attempt to maximize and to evaluate their reliability. Reliability is concerned with random measurement error. For instance, it is known that single-item (or single question) measures in questionnaires tend to be highly unreliable [45]. Therefore, when measuring complex concepts such as maturity or success, one is strongly advised to develop multiple-item measures (where more than one question is used to measure the concept) when possible. Furthermore, minimal evaluations of the reliability of measurement should be performed. Some procedures for doing so have been introduced in [17]. Reliability evaluation for measures of process implementation is important because, according to current evidence, process assessments are not perfectly reliable (e.g., see [22][23]).

The second problem concerns using coarse measures. For example, while the various dimensions of maturity have different effects on the process outcome, when combined into one dimension the overall effect may mask the dimensions that relate weakly to process outcomes, or vice versa. For instance, in one study [17] when four dimensions of maturity were summed up into one overall maturity dimension, the relationship between it and the quality of RE service was 0.33, which is statistically significant even though three of the dimensions are not individually related to RE service quality (i.e., the relationship was very small and not statistically significant). Therefore, this masking effect of coarse measures of process implementation distorts the effects of process implementation. Extreme caution should be taken when interpreting results from studies using coarse measures of process implementation. More reliable results would be obtained by considering individual dimensions separately.

Method of Data Analysis

The manner in which data is analyzed can have a substantial impact on the results. In particular, the analysis method should match the unit of analysis that we want to draw conclusions about. For example, if we want to draw conclusions about the benefits to organizations from implementing software process management practices, then it is necessary to conduct an analysis where the unit of analysis is the organization.

To illustrate this point, we consider the study reported in [26] and cited in [39] which found the relationships between CMM maturity levels and various measures of effectiveness to be statistically significant at the 0.05 alpha level. The reported data analysis pooled responses from 138 individuals representing 56 organizations (i.e., in many cases there were more than one respondent for each organization). This pooling of data effectively makes the analysis at the *individual* unit of analysis rather than the organization unit of analysis. This means that it is not appropriate to draw conclusions about the benefits of organizational maturity using this analysis approach. Another effect of this pooling is that it artificially increases the power of statistical tests and so artificially increases the likelihood of finding statistically significant relationships. When the observations are not pooled and one response per organization is used, the relationships are not statistically significant any more. Therefore, the strong conclusions drawn are not adequately supported by the results of the analyses. Of course, there are other ways of looking at this data. For example, all of the relationships were in the expected direction, and this is highly unlikely to occur by chance. Of course, any approach to formally test this hypothesis has to consider that the samples used for evaluating each relationship are not independent, and therefore tests like the sign test that assume independence would not be appropriate. But this conclusion is markedly different from the original one. Therefore, these results are not as compelling as would originally seem, and so data analysis choices should be critically examined.

Concluding Remarks

In this report we have presented data that can be used to plan and manage software process improvement efforts. The costs and benefits data pertain to different approaches to process improvement. The reader could identify the approach that is most relevant to his/her environment and use this data as guidance.

The accumulation of empirical evidence can also give us some useful lessons to increase the chances of attaining the potential benefits of SPI. We have discussed a number of substantive issues that have been identified by empirical results thus far. These issues should at least be considered during an SPI effort.

We have also attempted to shed some light on the methodological issues pertinent to studies that evaluate the benefits of software process improvement. It is clear that further research on this topic is forthcoming, and thus one should be careful in interpreting the results of these studies. Also, the methodological weaknesses that we have identified should serve as a challenge to future empirical researchers.

Acknowledgments

We wish to thank John Daly for reviewing an earlier version of this report..

References

[1] J. Bach: “Enough About Process: What we Need Are Heroes” In *IEEE Software,* 12(2):96-98, February 1995.

[2] V. Basili, G. Caldiera, F. McGarry, R. Pajerski, G. Page, and S. Waligora: “The Software Engineering Laboratory - An Operational Software Experience Factory” .In *Proceedings of the 14th International Conference on Software Engineering*, 1992.

[3] S. Benno and D. Frailey: “Software Process Improvement at DSEG: 1989-1995”. In *Texas Instruments Technical Journal,* 12(2):20-28, March-April 1995.

[4] L. Briand, C. Differding, and H. D. Rombach: “Practical Guidelines for Measurement-Based Process Improvement”. In *Proceedings of ISCN’96*, 1996.

[5] J. Brodman and D. Johnson: “What Small Businesses and Small Organizations Say about the CMM”. In *Proceedings of the 16th International Conference on Software Engineering*, pages 331-340, 1994.

[6] J. Brodman and D. Johnson: “Return on Investment (ROI) from Software Process Improvement as Measured by US Industry”. In *Software Process: Improvement and Practice,* Pilot Issue, John Wiley & Sons, 1995.

[7] J. Brodman and D. Johnson: “Return on Investment from Software Process Improvement as Measured by U.S. Industry”. In *Crosstalk,* 9(4):23-29, April 1996.

[8] C. Buchman: “Software Process Improvement at AlliedSignal Aerospace”. In *Proceedings of the 29th Annual Hawaii International Conference on Systems Science, Vol. 1: Software Technology and Architecture*, pages 673-680, 1996.

[9] K. Butler: "The economic Benefits of Software Process Improvement". In *Crosstalk*, 8(7):14-17, July 1995.

[10] D. Card: "Understanding Process Imporvement". In *IEEE Software*, pages 102-103, July 1991.

[11] B. Curtis, H. Krasner, and N. Iscoe: "A Field Study of the Software Design Process for Large Systems". In *Communications of the ACM*, 31(11):1268-1286, November 1988.

[12] M. Daskalantonakis: "A Practical View of Software Measurement and Implementation Experiences within Motorola". In *IEEE Transactions on Software Engineering*, pages 998-1010, November 1992.

[13] C. Deephouse, D. Goldenson, M. Kellner, and T. Mukhopadhyay: "The Effects of Software Processes on Meeting Targets and Quality". In *Proceedings of the Hawaiian International Conference on Systems Sciences*, vol. 4, pages 710-719, January 1995.

[14] R. Dion: "Process Improvement and the Corporate Balance Sheet". In *IEEE Software*, pages 28-35, July 1993.

[15] D. Drehmer and S. Dekleva: "Measuring Software Engineering Maturity: A Rasch Calibration". In *Proceedings of the International Conference on Information Systems*, pages 191-202, 1993.

[16] K. El Emam and N. H. Madhavji: "A Method for Instrumenting Software Evolution Processes and An Example Application". In *Notes From The International Workshop on Software Evolution, Processes, and Measurements*, Technical Report #94-04 NT, Software Engineering Test Lab, Department of Computer Science, University of Idaho, 1994.

[17] K. El Emam and N. H. Madhavji: "The Reliability of Measuring Organizational Maturity". In *Software Process: Improvement and Practice*, 1(1):3-25, 1995.

[18] K. El Emam and N. H. Madhavji: "Measuring the Success of Requirements Engineering Processes". In *Proceedings of the Second IEEE International Symposium on Requirements Engineering*, pages 204-211, 1995.

[19] K. El Emam and N. H. Madhavji: "A Field Study of Requirements Engineering Practices in Information Systems Development". In *Proceedings of the Second IEEE International Symposium on Requirements Engineering*, pages 68-80, 1995.

[20] K. El Emam, S. Quintin, and N. H. Madhavji: "User participation in the requirements engineering process: An empirical study". In *Requirements Engineering Journal*, 1:4-26, Springer-Verlag, 1996.

[21] K. El Emam and N. H. Madhavji: "The Impact of User Capability on Requirements Engineering Success". Submitted for publication, 1996.

[22] K. El Emam, D. R. Goldenson, L. Briand, and P. Marshall: "Interrater Agreement in SPICE-Based Assessments: Some Preliminary Results". In *Proceedings of the 4th International Conference on the Software Process*, 1996.

[23] K. El Emam, L. Briand, and R. Smith: “Assessor Agreement in Rating SPICE Processes”. Technical Report ISERN-96-09, International Software Engineering Research Network, 1996.

[24] P. Finlay and A. Mitchell: “Perceptions of the Benefits from the Introduction of CASE: An Empirical Study”. In *MIS Quarterly*, pages 353-370, December 1994.

[25] D. R. Goldenson: “Software Process Needs Analysis: Customer Perspectives”. Presentation Slides, Software Engineering Institute, June 1994.

[26] D. R. Goldenson and J. D. Herbsleb: “After the Appraisal: A Systematic Survey of Process Improvement, its Benefits, and Factors that Influence Success”. Technical Report, CMU/SEI-95-TR-009, Software Engineering Institute, 1995.

[27] R. Grady and D. Caswell: *Software Metrics: Establishing a Company-wide Program*. Prentice Hall, 1987.

[28] W. Hayes and D. Zubrow: “Moving On Up: Data and Experience Doing CMM-Based Process Improvement”. Technical Report CMU/SEI-95-TR-008, Software Engineering Institute, 1995.

[29] J. Herbsleb, A. Carleton, J. Rozum, J. Siegel, and D. Zubrow: “Benefits of CMM-Based Software Process Improvement: Initial Results”. Technical Report, CMU-SEI-94-TR-13, Software Engineering Institute, 1994.

[30] W. Humphrey, T. Snyder, and R. Willis: “Software Process Improvement at Hughes Aircraft”. In *IEEE Software*, pages 11-23, July 1991.

[31] W. Humphrey and W. Sweet: “A Method for Assessing the Software Engineering Capability of Contractors”. Technical Report CMU/SEI-87-TR-23, Software Engineering Institute, 1987.

[32] C. Jones: “The Pragmatics of Software Process Improvements”. In *Software Process Newsletter*, IEEE Computer Society TCSE, No. 5, pages 1-4, Winter 1996.

[33] C. Jones: *Assessment and Control of Software Risks*, Prentice-Hall, 1994.

[34] D. Kitson and S. Masters: “An Analysis of SEI Software Process Assessment Results: 1987-1991”. In *Proceedings of the International Conference on Software Engineering*, pages 68-77, 1993.

[35] P. Lawlis, R. Flowe, and J. Thordahl: “A Correlational Study of the CMM and Software Development Performance”. In *Software Process Newsletter*, IEEE TCSE, No. 7, pages 1-5, Fall 1996.

[36] L. Lebsanft: “Bootstrap: Experiences with Europe’s Software Process Assessment and Improvement Method”. In *Software Process Newsletter*, IEEE TCSE, No. 5, pages 6-10, Winter 1996.

[37] J. Lee and S. Kim: “The Relationship between Procedural Formalization in MIS Development and MIS Success”. In *Information and Management*, 22:89-111, 1992.

[38] W. Lipke and K. Butler: “Software Process Improvement: A Success Story”. In *Crosstalk*, 5(9):29-39, September 1992.

[39] B. Peterson: "Software Process Improvement Trends". In *Proceedings of the European Software Engineering Process Group Conference*, 1996.

[40] S. Rahhal: *An Effort Estimation Model for Implementing ISO 9001 in Software Organizations*. Master's Thesis, School of Computer Science, McGill University, October 1995.

[41] J. Rooijmans, H. Aerts, and M. van Genuchten: "Software Quality in consumer Electronics Products". In *IEEE Software*, pages 55-64, January 1996.

[42] H. Rubin: "Software Process Maturity: Measuring its Impact on Productivity and Quality". In *Proceedings of the International Conference on Software Engineering*, pages 468-476, 1993.

[43] Staff: "A Survey of the Surveys on the Benefits of ISO 9000". In *Software Process, Quality & ISO 9000*, 3(11):1-5, November 1994.

[44] H. Wohlwend and S. Rosenbaum: "Software Improvements in an International Company". In *Proceedings of the International Conference on Software Engineering*, pages 212-220, 1993.

[45] R. Zeller and E, Carmines: *Measurement in the Social Sciences: The Link Between Theory and Data*, Cambridge University Press, 1980.