Empirical Software Engineering, 5, 357–390, 2000.

An Instrument for Measuring the Key Factors of Success in Software Process Improvement

TORE DYBÅ* tore.dyba@informatics.sintef.no
SINTEF Telecom and Informatics

Abstract. Understanding how to implement SPI successfully is arguably the most challenging issue facing the SPI field today. The SPI literature contains many case studies of successful companies and descriptions of their SPI programs. However, there has been no systematic attempt to synthesize and organize the prescriptions offered. The research efforts to date are limited and inconclusive and without adequate theoretical and psychometric justification.

This paper provides a synthesis of prescriptions for successful quality management and process improvement found from an extensive review of the quality management, organizational learning, and software process improvement literature. The literature review was confirmed by empirical studies among both researchers and practitioners. The main result is an instrument for measuring the key factors of success in SPI based on data collected from 120 software organizations. The measures were found to have satisfactory psychometric properties. Hence, managers can use the instrument to guide SPI activities in their respective organizations and researchers can use it to build models to relate the facilitating factors to both learning processes and SPI outcomes.

Keywords: Software process improvement, success factors, measurement instrument.

Introduction

During the last decade, the software industry has been more and more concerned about software process improvement (SPI). Consequently, we have witnessed a proliferation of models and initiatives all claiming to increase the likelihood of succeeding with SPI initiatives. The SPI literature is full of case studies and anecdotal evidence of successful companies and descriptions of their SPI programs. Several authors repeatedly discuss the importance of certain critical success factors. However, there has been no systematic attempt to synthesize and organize the prescriptions offered. The research efforts to date are limited and inconclusive and without adequate theoretical and psychometric justification. Even for commonly recognized factors such as management commitment and employee participation, no operational measures are available.

Despite an increasing focus on the nature and value of empirical software engineering (e.g. Basili *et al.*, 1986; Basili and Selby, 1991; Basili, 1996; Harrison *et al.*, 1999; Jeffery and Votta, 1999), poor theory development and inadequate measurement of constructs seems to be the norm for most studies regarding the key factors of success in SPI. The data collection device (in this study a questionnaire) used for measurement is commonly referred to as an *instrument* (Davis, 1996). However, instrument validation and reliability issues have been inadequately addressed in software engineering research, with only a handful of researchers

* SINTEF is The Foundation for Scientific and Industrial Research at the Norwegian Institute of Technology. Mail address: N-7465 Trondheim, Norway. Tel +47 73 59 29 47, Fax +47 73 59 29 77

(e.g. El Emam and Madhavji, 1995, 1996; Goldenson *et al.*, 1999; Fusaro *et al.*, 1998; El Emam, 1998; El Emam and Birk, 2000) devoting serious attention to these issues.

Measurement instruments in both research and practice are expected to be valid and reliable (Straub, 1989; Anastasi and Urbina, 1997). The basic point is that users of a given instrument should obtain similar results. Thus, psychologists and psychometric theorists have developed rigorous methods for constructing reliable and valid instruments to measure variables in the social sciences (e.g. Cronbach, 1951; Likert, 1967; Nunnally, 1978; Nunnally and Bernstein, 1994; Anastasi and Urbina, 1997).

Hence, for the present research we define *instrument construction* as:

> the process of developing the data collection device in order to define and obtain relevant data for a given research question.

There are many reasons why researchers in SPI should pay closer attention to instrumentation. First, concerns about instrumentation are closely connected with concerns about rigor in empirical research methods for SPI. Second, greater attention to instrumentation permits confirmatory, follow-up research to use a tested instrument, hence, promoting cooperative research efforts (Hunter and Schmidt, 1990). Third, closer attention to instrumentation brings greater clarity to the formulation and interpretation of research questions (Straub, 1989). Finally, lack of validated measures in confirmatory research raises serious questions about the trustworthiness of the findings of the study.

In SPI research, as in social research and business research, we attempt to understand real-world phenomena through expressed relationships between research constructs (Blalock, 1969). However, these constructs are neither directly measurable nor observable, but are believed to be latent in the phenomenon under investigation. In SPI, for example, casual constructs like "organizational maturity" is thought to influence outcome constructs like "organizational performance." Neither of these constructs is directly observable, but relevant measures can be operationalized to serve as surrogates for these constructs.

Measurement of research constructs is neither simple nor straightforward. However, instrumentation techniques are available that allow us to construct research instruments that constitutes acceptable levels of reliability and validity. The process of developing the research instrument for this study was based on generally accepted psychometric principles of instrument design, and was carried out according to the nine steps shown in Figure 1.

The sections in this paper are arranged according to these steps. Section 1 presents the literature review. Section 2 presents the process of identifying the key factors of SPI success. The identification of items for each factor is described in section 3, and the construction of measurement scales is presented in section 4. In section 5, we present a pilot test of the measurement scales and the overall instrument before data was collected (section 6). Finally, sections 7, 8, and 9 are concerned with the reliability of the measurement scales, detailed item analysis, and validity of the measurement scales, respectively.

1. Literature Review

Software process improvement has its roots in quality management and it is closely related to "second generation" (French and Bell, 1999) organizational development approaches,

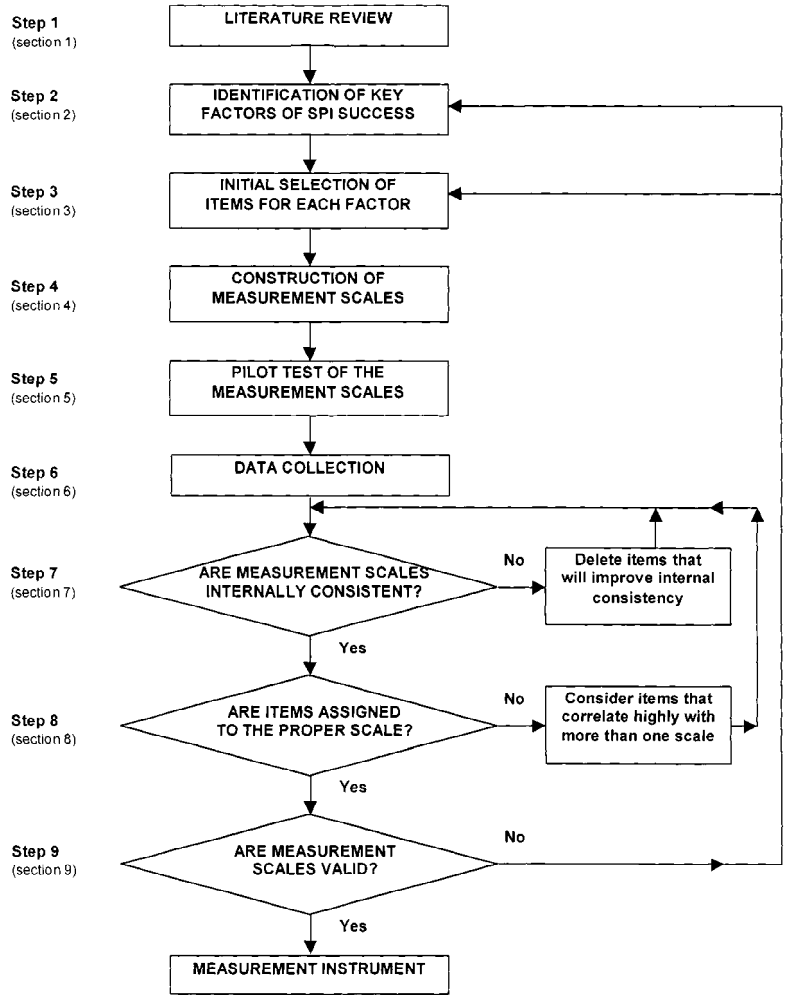

Figure 1. Instrument development process (adapted from Saraph *et al.*, 1989).

specifically to organizational learning. Hence, this section reports from an extensive review of the quality management, organizational learning and software process improvement literature regarding key factors for success. Nearly 200 prescriptions for success were derived from the literature.

Table 1. Perspectives on TQM.

Deming's 14 principles[1]		The Juran Trilogy[2]		Crosby's 14 quality steps[3]	
1.	Constancy of purpose	I.	*Quality Planning*	1.	Management commitment
2.	Adopt the new philosophy		Establish quality goals	2.	Quality improvement teams
3.	Cease dependence on inspection		Identify customers and their needs	3.	Quality measurement
4.	Don't award business on price		Develop products and processes	4.	Cost of quality evaluation
5.	Constant improvement	II.	*Quality Control*	5.	Quality awareness
6.	Institute training on the job		Evaluate performance	6.	Corrective action
7.	Institute leadership		Compare to goals and acts	7.	Zero-defects committee
8.	Drive out fear	III.	*Quality Improvement*	8.	Supervisor training
9.	Break down barriers		Establish infrastructure	9.	Zero-defects day
10.	Eliminate slogans		Identify improvement projects and teams	10.	Goal-setting
11.	Eliminate work standards		Provide resources and training	11.	Error cause removal
12.	Pride of workmanship		Establish controls	12.	Recognition
13.	Education and retraining			13.	Quality councils
15.	Take action			14.	Do it over again

Sources: [1]Deming (1986), [2]Juran (1992), [3]Crosby (1979).

1.1. Critical Factors of Quality Management

Understanding the "quality revolution" is an important prerequisite for understanding software process improvement. However, quality management is often obscured by both confusion and misunderstanding. This potential for misunderstanding is partly related to quasi-religious and sectarian controversies, and partly related to the most important feature of modern quality management: it directs attention to the improvement of production processes, and not simply to the characteristics of the product (Winter, 1994). In this respect, modern quality management opposes the original quality control principles of assuring that the characteristics of the end product fall within preassigned tolerance limits.

The current state-of-the-art in quality management has more than anything else been shaped by quality gurus such as William Edwards Deming (1982, 1986), Joseph M. Juran (1992, 1999), Philip Crosby (1979, 1984, 1996), and their quality frameworks. Table 1 summarizes Deming's 14 principles, the Juran trilogy, and Crosby's 14 quality steps. These and other authors (e.g. Ahire *et al.*, 1996; Black and Porter, 1996; Feigenbaum, 1991; Garvin 1983, 1984; Ishikawa, 1986, 1990; Powell, 1995; Saraph *et al.*, 1989; Taguchi, 1986; Taguchi *et al.*, 1989; Yusof and Aspinwall, 1999) repeatedly discuss the importance of critical factors such as leadership involvement, employee participation, measurement, and process management to improve the quality performance in organizations.

The recent revision of the ISO 9000 family of standards distances itself from the term "quality assurance," encouraging the adoption of a process approach to "quality management." In this revision, eight quality management principles (see Table 2) to facilitate the achievement of quality objectives are identified (ISO/DIS 9000, 2000). Furthermore, each of the major clauses in ISO 9004, Quality Management Systems—Guidelines for Performance Improvements, is based on these principles (ISO/DIS 9004, 2000).

Table 2. ISO 9000 quality management principles (ISO/DIS 9000, 2000).

ISO 9000 Quality Management Principles	
a) *Customer focus:* organizations depend on their customers and therefore should understand current and future customer needs, should meet customer requirements and should strive to exceed customer expectations.	e) *System approach to management:* identifying, understanding and managing a system of interrelated processes for a given objective contributes to the effectiveness and efficiency of the organization.
b) *Leadership:* leaders establish unity of purpose, direction, and the internal environment of the organization. They create the environment in which people can become fully involved in achieving the organization's objectives.	f) *Continual improvement:* a permanent objective of the organization is continual improvement.
c) *Involvement of people:* people at all levels are the essence of an organization and their full involvement enables their abilities to be used for the organization's maximum benefit.	g) *Factual approach to decision making:* effective decisions are based on the logical or intuitive analysis of data and information.
d) *Process approach:* a desired result is achieved more efficiently when related resources and activities are managed as a process.	h) *Mutually beneficial supplier relationships:* the ability of the organization and its suppliers to create value is enhanced by mutually beneficial relationships.

No single model has yet established itself as a basis for Total Quality Management (TQM) theory. As a result, there has been a trend in organizations to use TQM frameworks based upon the assessment criteria from key quality awards such as the Deming Prize in Japan, the Malcolm Baldridge National Quality Award in the United States and the European Quality Award. Founded in 1988, The European Foundation for Quality Management (EFQM) has developed the EFQM Excellence Model, which is the basis for the European Quality Award. The EFQM Excellence Model is based on nine criteria (see Table 3), of which five are "enablers" and four are "results" (EFQM, 1999).

Despite TQM's apparent widespread dissemination, and the claims by adherents that any firm can imitate TQM, there are powerful reasons to believe that TQM is imperfectly imitable (Powell, 1995). Rogers (1995) has shown that the rate of adoption is determined by the characteristics of an innovation as *perceived by the potential adopter*, and not whether it has produced any advantages for competitors. His research has shown that the diffusion of innovations depend on the following five factors (Rogers, 1995): (1) *relative advantage*—the degree to which an innovation is perceived as being better than the idea it supersedes; (2) *compatibility*—the degree to which an innovation is perceived as consistent with the existing values, past experience, and needs of potential adopters; (3) *complexity*—the degree to which an innovation is perceived as relatively difficult to understand and use; (4) *trialability*—the degree to which an innovation may be experimented with on a limited basis; and (5) *observability*—the degree to which the results of an innovation are visible to others.

Table 3. EFQM Excellence Model criteria and sub-criteria (EFQM 1999).

The 1999 EFQM Excellence Model Criteria	
1. *Leadership*	
1a)	Leaders develop the mission, vision and values and are role models of a culture of Excellence.
1b)	Leaders are personally involved in ensuring the organization's management system is developed, implemented and continuously improved.
1c)	Leaders are involved with customers, partners and representatives of society.
1d)	Leaders motivate, support and recognize the organization's people.
2. *Policy and Strategy*	
2a)	Policy and strategy are based on the present and future needs and expectations of stakeholders.
2b)	Policy and strategy are based on information from performance measurement, research, learning and creativity related activities.
2c)	Policy and strategy are developed, reviewed and updated.
2d)	Policy and strategy are deployed through a framework of key processes.
2e)	Policy and strategy are communicated
3. *People*	
3a)	People resources are planned, managed and improved.
3b)	People's knowledge and competencies are identified, developed and sustained.
3c)	People are involved and empowered.
3d)	People and the organization have a dialogue.
3e)	People are rewarded, recognized and cared for.
4. *Partnerships and Resources*	
4a)	External partnerships are managed.
4b)	Finances are managed.
4c)	Buildings, equipment and materials are managed.
4d)	Technology is managed.
4e)	Information and knowledge are managed.
5. *Processes*	
5a)	Processes are systematically designed and managed.
5b)	Processes are improved, as needed, using innovation in order to fully satisfy and generate increasing value for customers and other stakeholders.
5c)	Products and services are designed and developed based on customer needs and expectations.
5d)	Products and services are produced, delivered and serviced.
5e)	Customer relationships are managed and enhanced.
6. *Customer Results*	
6a)	Perception measures.
6b)	Performance indicators.
7. *People Results*	
7a)	Perception measures.
7b)	Performance indicators.
8. *Society Results*	
8a)	Perception measures.
8b)	Performance indicators.
9. *Key Performance Results*	
9a)	Key performance outcomes.
9b)	Key performance indicators.

Furthermore, diffusion of innovation models emphasize the importance of *homophily*, which Rogers (1995) defined as the degree to which the innovator and the potential adopter are similar in certain attributes such as objectives, beliefs, norms, experience, and culture. Heterophily is the opposite of homophily, and, according to Rogers (1995), "*one of the most distinctive problems in the diffusion of innovations is that the participants are usually quite heterophilous.*" (p. 19, italics in original). Hence, vital difference between innovators and potential adopters act as key barriers to imitation (Powell, 1995).

Schaffer and Thomson (1992) argued that most "activity centered" change programs, such as TQM, are based on a fundamentally flawed logic that confuses ends with means, and

processes for outcomes. They claimed that there are six reasons why quality improvement programs fail: (1) process (rather than results) orientation; (2) too large scale and diffused; (3) bad results are excused for the sake of program success; (4) delusional measurements of success; (5) staff- and consultant-driven; and (6) bias to orthodoxy, not cause and effect.

In contrast to the activity centered programs, Schaffer and Thomson (1992) argued for a "results-driven" approach to improvement. The results-driven approach, they claimed, have four key benefits that activity-centered programs generally miss: (1) they introduce managerial and process innovations only as they are needed; (2) they are based on empirical testing that reveals what works; (3) they are focused on short-term, incremental projects that quickly yield tangible results; and finally (4) management creates a continuos learning process by building on the lessons of previous phases in designing the next phase of the program.

Even though we have witnessed a widespread research and industrial use of TQM, there is still no general agreement on a set of operational measures of quality management in terms of critical factors for success. Only a handful of researchers have made contributions towards a scientific methodology for diagnostic assessments of TQM programs (e.g. Garvin, 1983, 1984; Saraph *et al.*, 1989; Powell, 1995; Ahire *et al.*, 1996; Black and Porter, 1996).

Garvin (1983, 1984) conducted early empirical studies of quality practices and their impact on quality performance. He developed a set of critical factors based on systematic on-site observations and comparisons of quality management practices between manufacturers in the United States and Japan. These studies concluded that the quality leaders performed especially well in several areas of quality management. They had strong management support for quality, a comprehensive goal-setting process, and a high level of cross-functional participation.

Saraph, Benson and Schroeder (1989) developed a set of eight critical factors of quality management based on the quality literature. They pioneered the development of valid and reliable instruments, based on accepted psychometric principles, to measure quality management practices. The following factors were included in their final instrument: Role of management leadership and quality policy, role of quality department, training, product/service design, supplier quality management, process management, quality data and reporting, and employee relations.

Black and Porter (1996) identified ten critical factors of TQM based on an empirical study among 204 managers with membership in the EFQM. The Baldridge Award model was taken as the basis for their study. However, the empirically derived factors did not match the categories suggested by the Baldridge model. Also, the factor analytic approach used in this study resulted in several highly complex factors, which makes them difficult to interpret. In practical terms, however, Black and Porter's (1996) findings can be used to improve existing self-assessment frameworks such as the Baldridge and European Quality Award criteria.

Through a detailed analysis of the literature, Ahire, Golhar and Waller (1996) identified, and empirically validated, 12 implementation constructs of integrated quality management strategies. However, in an empirical study of TQM as competitive advantage, Powell (1995) claimed that most features generally associated with TQM (e.g. quality training, process improvement, and benchmarking) do not generally produce advantage. On the

other hand, he found that features such as open culture, employee empowerment, and executive commitment, could produce advantage.

1.2. Critical Factors of Organizational Learning

Studies of organizational learning has more than anything else been shaped by the works of Argyris and Schön (1978, 1996). Furthermore, Senge (1990; Senge *et al.*, 1994, 1999) has popularized the concept of the "learning organization," Nonaka (1991, 1994; Nonaka and Takeuchi, 1995) has emphasized the "knowledge creating company," Choo (1995, 1998) has developed the notion of the "knowing organization," Hunt and Buzan (1999) argued for creating a "thinking organization," and finally March (1999) has highlighted the pursuit of "organizational intelligence."

Argyris and Schön made an important distinction between the concepts of *single-loop* and *double-loop* learning, what Senge (1990) has called the difference between adaptive and generative learning, or what others have called the difference between (1) maintenance and growth vs. (2) transformation. Similarly, March (1991, 1999) makes a distinction between the exploitation of existing knowledge and the exploration of new knowledge. Common to all of these constructs, is that the lower level involves improving existing behaviors and making progress toward stated goals, while the higher level requires questioning the appropriateness of the goals, and recognizing the subjectivity of meaning.

Evolutionary learning strategies represent relatively small or incremental changes, in the organization's products, procedures, or services. They are new to the organization but reflect an adaptation or simple adjustment of existing practices, and their implementation rarely requires changes in organizational structures or processes. In contrast, radical learning strategies represent larger changes in organizational products, procedures, or services. They reflect broader shifts in perspective and reorientation of existing practices and often require major changes in organizational structures or processes to implement.

This distinction is also consistent with a characterization of learning strategies ranging from adaptive to innovative. Organizations with adaptive styles work within existing structures to make incremental changes and "do things better." In contrast, organizations with innovative styles treat current structures as part of the problem and make more radical changes by "doing things differently." In other words, software organizations can engage in two broad kinds of learning strategies. They can engage in *exploitation*—the adoption and use of existing knowledge and experience, and they can engage in *exploration*—the search for new knowledge, either through imitation or innovation (Dybå, 2000a).

Nonaka (1991, 1994) discussed five conditions required at the organizational level to promote a knowledge spiral of these learning strategies, Senge (1990) proposed five disciplines for creating a learning organization, and Garvin (1993) emphasized five main activities that learning organizations are skilled at. Table 4 summarizes these prescriptions.

Furthermore, in an empirical study by Nevis, DiBella and Gould (1995), ten facilitating factors that expedite learning were identified: (1) Scanning imperative, (2) performance gap, (3) concern for measurement, (4) experimental mind-set, (5) climate of openness, (6) continuous education, (7) operational variety, (8) multiple advocates, (9) involved leadership, and (10) systems perspective.

Table 4. Facilitating factors identified in organizational learning.

Nonaka's enabling factors[1]	**Senge's five disciplines**[2]	**Garvin's building blocks**[3]
• Intention	• Systems Thinking	• Systematic problem solving
• Autonomy	• Personal Mastery	• Experimentation
• Fluctuation and Creative Chaos	• Mental Models	• Learning from past experience
• Redundancy	• Shared Vision	• Learning from others
• Requisite Variety	• Team Learning	• Transferring knowledge

Sources: [1]Nonaka (1994), [2]Senge (1990), [3]Garvin (1993).

Table 5. Facilitating factors identified in software process improvement.

Humphrey's six principles[1]	**Zahran's 10 CSFs**[2]	**Basili's paradigm**[3]
• Major changes to the software process must start at the top	1. Alignment with the business strategy and goals	• Acquisition of core competencies through (1) a control cycle and (2) a capitalization cycle
• Ultimately, everyone must be involved	2. Consensus and buy-in from all stakeholders	• Goal-oriented measurement
• Effective change is built on knowledge	3. Management support	• Experience reuse and organizational sharing
• Change is continuous	4. Dedicated resources	
• Software process changes won't stick by themselves	5. Sensitivity to the organizational context	
• Software process improvement requires investment	6. Management of change	
	7. Prioritization of actions	
	8. Support infrastructure	
	9. Monitoring the results of SPI	
	10. Learning from the feedback results	

Sources: [1]Humphrey (1989), [2]Zahran (1998), [3]Basili and Caldiera (1995).

1.3. Critical Factors of Software Process Improvement

Watts Humphrey (1989, 1997) and Victor R. Basili (e.g. Basili and Rombach, 1988) have been the pioneers and leaders in the field of software process improvement. Humphrey (1989) identified six basic principles of software process change. Zahran (1998) proposed ten critical factors for successful implementation of software process improvement. Basili and Caldiera (1995) focus on reuse of experience and learning by using the quality improvement paradigm (QIP) for developing core competencies, and by supporting the QIP process with goal-oriented measurement (GQM) and an organizational infrastructure (EF). Table 5 summarizes these concepts.

Goldenson and Herbsleb (1995) conducted a survey of 138 individuals from 56 organizations in the United States and Canada to evaluate various organizational factors that were believed to promote or hinder successful SPI after a CMM-based assessment. The factors that were found to be statistically significant in their study are summarized in Table 6. El Emam *et al.* (1998) made a reanalysis of Goldenson and Herbsleb's (1995) study, using multivariate analysis instead of the simple statistical analytic methods used in the initial

Table 6. Facilitating factors and barriers to SPI (Goldenson and Herbsleb, 1995).

Organizational Factors	Barriers
• Senior management monitoring of SPI • Compensated SPI responsibilities • SPI goals well understood • Technical staff involved in SPI • SPI people well respected • Staff time/resources dedicated to process improvement	• Discouragement about SPI prospects • SPI gets in the way of "real" work • "Turf guarding" inhibits SPI • Existence of organizational politics • Assessment recommendations too ambitious • Need guidance about how to improve • Need more mentoring and assistance

report. Based on this reanalysis, they identified focused SPI effort, commitment to SPI, politics, respect, and turnover as the key factors.

Within the SPICE Trials, a similar study to that of Goldenson and Herbsleb (1995) was conducted by El Emam, Fusaro and Smith (1999) with 18 organizations in Europe, Canada and Australia that had performed assessments using the ISO/IEC 15504 standard for software process assessment. In their study, three types of independent variables were tested: "organizational factors," "process factors" and "barriers." Results of the bivariate relationship analysis showed that none of the identified barriers were related to success in addressing the findings from an assessment. Of the organizational factors, only "SPI goals being well understood" and "Technical Staff involvement in SPI" were found to be critical for addressing the findings from an assessment. Finally, only one process factor, "Creating process action teams," was found to be statistically significant in addressing the assessment findings.

Stelzer, Mellis and Herswurm (1996) identified the following key success factors in their study of software process improvement via ISO 9000: (1) Definition and documentation of the status quo, (2) identification of best practices, (3) identification of business processes, (4) simplification of routine procedures, (5) internal audits, (6) impetus and incentive, (7) team spirit, (8) workshop and regular meetings, (9) definition of a common language, and (10) customer perception surveys.

Furthermore, Stelzer and Mellis (1998) analyzed published experience reports and case studies of 56 software organizations that had implemented an ISO 9000 quality system or that had conducted a CMM-based SPI initiative. The result of this meta-analysis was a set of ten factors that affect organizational change in SPI. In rank order, these factors were: (1) Management commitment and support, (2) staff involvement, (3) providing enhanced understanding, (4) tailoring improvement initiatives, (5) managing the improvement project, (6) change agents and opinion leaders, (7) stabilizing changed processes, (8) encouraging communication and collaboration, (9) setting relevant and realistic objectives, and (10) unfreezing the organization.

Moreover, in a survey of 87 projects from different organizations, Deephouse *et al.* (1996) assessed the effectiveness of software processes on project performance. The results from this study showed that certain practices, such as project planning and cross-functional teams, were consistently associated with favorable outcomes, while other practices such as process

training, stable environment, user contact, design reviews, and prototyping had little impact on project outcomes.

Finally, ISO/IEC 15504-7 (1998) highlight cultural and management issues as fundamental to succeed with software process improvement and organizational change. The standard argues that SPI should be strongly supported by leadership, communication and motivation throughout the whole organization; that improvement actions only can be carried out efficiently if the appropriate cultural issues are acknowledged and addressed at all levels; and finally, that major problems found in software processes often arise from cultural issues. Consequently, cultural issues should be one of the factors considered in prioritizing improvement actions.

In the rest of the paper the process of developing the instrument based on the literature-derived prescriptions and an extensive empirical research is described. Other sets of factors could be developed or different aspects of the factors could have been defined. However, this set seems to capture most of the important aspects of successful SPI as espoused by leading practitioners and researchers. This is also, as we shall see, supported by the results of reliability and validity analyses that show that the instrument has desirable psychometric properties.

2. Key Factors of SPI Success

The starting point for developing the instrument was the extensive literature review discussed in the preceding section. The above authors have emphasized slightly different sets of organizational requirements for successful quality management and organizational change based on their personal judgment and experience. With a few exceptions (e.g. Saraph *et al.*, 1989; Powell, 1995; Ahire *et al.*, 1996; Black and Porter, 1996; El Emam *et al.*, 1998) their requirements were not formulated on the basis of systematic empirical research.

As a complement to the literature study, we conducted an extensive exploratory study of factors enabling SPI success in four Norwegian software companies. The data collection method used to derive key factors for success was three questionnaire items and follow-up group interviews (feedback sessions) in each company. For the purpose of the results presented in this paper, data collection focused on answering one question:

> In your opinion, what are the three most important factors enabling SPI success in your organization?

In order to gain more background information about the key factors of SPI success, our questioning strategy included two additional questions. First we asked the subjects about their most important argument *in favor* of SPI in their organization. Second we asked the subjects about their most important argument *against* SPI in their organization. The answers to these questions gave us valuable information for interpreting the factors.

In total, 54 software managers, quality managers, software developers and customer representatives answered the three questions and participated in the subsequent feedback sessions. The outcome of this study was a set of five enabling factors for SPI success: management commitment, participation, learning from experience, goal-/business orientation and measurement. Each of these five factors was mentioned by at least three subjects.

Complementing the literature review and the exploratory study in the four software companies, we conducted a review process with eleven SPI experts in both academia and industry. The single most important criteria for choosing experts to this panel was "hands-on experience with SPI projects in the software industry." Thus, we included software managers, senior software engineers and researchers that actively participated *either* in the largest software process improvement effort to date in Norway, called SPIQ (Software Process Improvement for better Quality) (see Dybå, 2000b), *or* in a Process Improvement Experiment (PIE) within the European Systems and Software Initiative (ESSI).

The experts were, thus, chosen from companies of different sizes with a wide range of products and markets, developing either software or combined software and hardware products. Some of the industry expert's companies belonged to the traditional electronics based IT-industry, while others belonged to the new dot-com industry. Besides SINTEF, experts from academia were chosen from the Norwegian University of Science and Technology, and the University of Oslo, which made up the research partners in the SPIQ project. Of the panel of eleven experts seven had a Ph.D. in computer science, six had project management or SPI responsibility in industry, and five were researchers in software engineering and SPI in universities or research institutes.

The use of such an expert review process is common in instrument development, and the particular process used in this study was similar to that of El Emam and Madhavji (1996). The review process consisted of three consecutive activities: (1) refinement of the literature-derived prescriptions, (2) prioritization of the hypothesized categories, and (3) prioritization of indicators within each category.

Each expert received a memorandum with a brief summary of the prescriptions for success found from the literature review. Furthermore, as a result of this literature review, the experts were presented with an initial list of seven key factors for SPI success and a set of potential indicators for each factor based on the author's judgement. Nearly all of the authors supported each of these factors; hence, together they were judged to define essential aspects of SPI.

Each expert was asked to (1) review the findings from the literature and the initial list of key factors, (2) delete the factors considered to have secondary or no importance, (3) add new factors considered to have high importance, and (4) give each of the *remaining* factors a priority score. The experts were also asked to add any literature references that, according to their judgement, was not represented by the review, i.e. that would add new prescriptions not covered by the presented literature review. Subsequently, the list of factors was modified, refined, and reworded during the data collection.

A rank ordering was obtained using the total proportion of respondents who ranked a factor as critical for SPI success, as opposed to being of somewhat importance. During this step of the expert review process, it became clear that the experts placed similar levels of emphasis on the SPI success factors. The results of this survey and the subsequent rank ordering indicated which factors of SPI success were considered more important, and which were considered less important.

During this process, it was clear that *the five factors that were ranked highest by the eleven SPI experts were the same factors that resulted from the exploratory study among the 54 managers and developers in the four software companies*. The two lowest ranking factors

from the expert review process (experimentation and operational variety) were combined into one factor (exploration of new knowledge). Hence, the result of this investigation was the identification of six key facilitating factors for SPI success. These six factors are described in Table 7.

In the next section, we describe the identification of indicators for each of the facilitating factors of SPI success.

3. Item Creation

Using the prescriptions found in the literature, several representative indicators were defined for each factor. In total, 59 indicators were initially defined for the original set of seven facilitating factors. These indicators are presented in Appendix A.

A process similar to that of identifying the key factors of SPI success was used to define the key indicators for each factor. Hence, for each factor, the eleven experts were asked to (1) delete indicators of secondary or no importance, or indicators that they considered inappropriate as descriptions of the factor under consideration, (2) add new indicators according to their judgement, and (3) give each of the *remaining* indicators a priority score.

As for the factors, a rank ordering was now obtained for the indicators, based on the experts' judgement. During this step of the expert review process it became clear that different experts placed different levels of emphasis on the indicators. The results of the subsequent rank ordering suggested which indicators that were considered more important or less important. Furthermore, according to Bagozzi (1996): "for scales comprised of the sums of items, about 4 to 8 items are usually adequate" (p. 41). Thus, following the expert review, the five to eight highest ranked indicators for each factor were chosen, for a total of 38 indicators. These indicators are shown in rank order within each factor in Table 7.

4. Construction of Measurement Scales

Based on the results from the preceding step, we defined one question for each indicator such that the theoretical abstraction of each indicator could be related more closely to everyday work situations. Furthermore, a subjective rating scale accompanied each question.

In selecting the number of points on a rating scale, Guilford (1954) suggested several considerations. If too few scale points are used, the answer scale is obviously coarse, and much information is lost because the scale does not capture the discriminatory powers that respondents are capable of making. Conversely, by using too many scale points, the scale can become graded so finely that it is beyond the respondents' limited powers of discrimination. Indeed, Miller (1956) argued that the average individual can process seven, plus or minus two (the "magical number") chunks of information at a time.

Likert and Roslow (1934) investigated the reliability of attitude scales by using three variations of the Likert scale. In addition to the original 5-point scale (Likert, 1932), they also used a 3-point and a 7-point scale. They concluded that the 5-point scale consistently yielded higher reliabilities than either of the two other scales.

Table 7. Key factors of SPI success.

Key Factors of SPI Success	Indicators for the Key Factors of SPI Success
Business Orientation	• Extent of goal-orientation in SPI activities. • Extent to which SPI goals and policy are understood. • Degree of integrating SPI actions with "real" work. • Extent to which SPI goals are aligned with business goals. • Degree of balance between short-term and long-term goals.
Leadership Involvement	• Degree of management support to SPI activities. • Acceptance of responsibility for SPI by management. • Degree to which management considers SPI as a way to increase competitive advantage. • Degree of participation by management in the SPI process. • Amount of review of SPI issues in top management meetings.
Employee Participation	• Extent of employee involvement in decisions about what should best be done at their own level (co-determination). • Extent to which employees contribute with improvement proposals. • Extent of developer participation in the formalization of routines. • Extent of on-going dialogue and discussion about software development. • Extent to which employees have responsibility for SPI. • Extent of developer participation in SPI goal setting. • Extent of on-going dialogue and discussion about SPI. • Extent of employee participation in development planning*.
Concern for Measurement	• Importance of measuring performance. • Availability of quality data (defects, timeliness, etc). • Extent to which quality data is available to developers. • Extent to which quality data is available to managers. • Extent to which quality data is used in SPI work. • Amount of feedback provided to project teams on their performance.
Exploitation of existing knowledge (learning by experience)	• Extent to which existing knowledge is exploited. • Extent of learning from past experience. • Degree to which formal routines are based on past experience. • Degree of systemization of past experience. • Degree of internal experience transfer. • Extent of risk aversion*.
Exploration of new knowledge (learning by experimentation)	• Degree of adaptability to rapid change, increasing complexity and environmental uncertainty. • Extent to which innovation/creativity is encouraged. • Extent of variety in the use of methods, techniques and tools. • Degree of experimentation with new ideas, strategies, and technologies. • Ability to question underlying values. • Degree of flexibility in task execution. • Degree of detail in task specifications (minimum critical specification). • Importance of matching the variety and complexity of the organization's environment.

*These items were eventually deleted to improve the reliability of the instrument.

More recently, Lissitz and Green (1975) conducted a Monte Carlo study of the effects of the number of scale points and homogeneity upon reliability. Their study showed that there was an increase in reliability as the number of points increased from two to five. However, reliability leveled off beyond five scale points. Consequently, Lissitz and Green (1975) concluded that since respondents are fallible, even fewer than five scale points may be necessary.

Qualitative studies by Van de Ven and Ferry (1980) support the presented conclusions by Likert and Roslow (1934) and Lissitz and Green (1975). Guided by these studies, 5-point bipolar Likert scales were constructed for all questions in our final questionnaire. Responses were scored from 1 to 5, with a value of 1 indicating "Strongly disagree" and a value of 5 indicating "Strongly agree." Thus, we refer to each question and its associated 5-point scale as an *item*. A typical item is presented below:

	Strongly disagree **(1)**	Disagree **(2)**	Neither agree nor disagree **(3)**	Agree **(4)**	Strongly agree **(5)**
Management is actively supporting SPI activities.	□	□	□	□	□

Random errors in single items will sometimes inflate and sometimes deflate the observed estimate of the true score. Therefore, when repeated measurements from the same respondents are taken over time, there will be inconsistencies in the observations. Consequently, single-item measures in questionnaires tend to be highly unreliable (Nunnally and Bernstein, 1994; Spector, 1992). On the other hand, when multiple items are combined into an estimate of the true score, errors tend to average out, leaving a more accurate and consistent (reliable) measurement over time (Spector, 1992).

Hence, to reliably measure complex concepts such as SPI success and facilitating factors, we developed multiple-item scales where more than one question was used to measure each concept. The actual level of practice for each facilitating factor is, thus, represented by the sum of the item ratings for that factor. We refer to these sums as *scale scores*. A vector of the scale scores for the six factors can thus be used to predict the software organization's chances of success with its SPI program.

As an example of a scale (see Table 8) we present the scale developed for the facilitating factor "Leadership Involvement" based on the five indicators resulting from the expert review process (the complete validated instrument is included in Appendix B).

5. Pilot Test of Measurement Scales

The next step in the instrument development process was a pilot test of the measurement scales and of the overall instrument. In general, the pilot sample should be as similar to the target population as possible (Nunnally and Bernstein, 1994; Fink and Kosecoff, 1998). Since the primary objective of this research was to develop an instrument to measure a software organization's score on the six facilitating factors of SPI success, development managers and quality managers were considered appropriate subjects. Including both of

Table 8. Leadership involvement scale.

Leadership Involvement	Strongly disagree **(1)**	Disagree **(2)**	Neither agree nor disagree **(3)**	Agree **(4)**	Strongly agree **(5)**
Management is actively supporting SPI activities.	☐	☐	☐	☐	☐
Management accepts responsibility for SPI.	☐	☐	☐	☐	☐
Management considers SPI as a way to increase competitive advantage.	☐	☐	☐	☐	☐
Management is actively participating in SPI activities.	☐	☐	☐	☐	☐
SPI issues are often discussed in top management meetings.	☐	☐	☐	☐	☐

these management groups were also important in order to have a more balanced view, since their attitudes may differ widely.

Furthermore, since this was a pilot test, the sample size was kept quite small. Questionnaires were distributed to a convenient sample of twelve managers in eight companies that either participated in the SPIQ program or were involved in an ESSI PIE. Each manager assessed his/her company by rating each measurement item using the scales described in the previous section. The managers mailed the completed questionnaires directly to the author to ensure confidentiality and anonymity of each response.

The goals of the pilot test were twofold: (1) to ensure that the mechanisms of compiling the questionnaire had been adequate, and (2) to make an initial reliability assessment of the measurement scales. The first aim of the pilot test was accomplished by having two professors of SPI research and one professor in psychological testing review the completed questionnaire, and comment on its length, wording, instructions, and format before it was sent out. Subsequently, each manager was asked to make the same kind of comments. Non of the comments from this review implied a need to change the questionnaire.

Analyzing the correlation of items within each scale (item-item), the corrected item-to-total (item-scale) correlations, and the item standard deviation scores accomplished the second aim of the pilot test. Based on this analysis, we deleted one item in the "Employee participation" scale (marked with an asterisk in Table 7). This item was also the lowest ranked item in the previously described expert review of the "Employee participation" scale.

6. Subjects and Instrument Administration

Target respondents for the study were software managers and quality managers in Norwegian IT companies developing software. These managers were chosen since they serve

as a bridge between the visions of top management and the often chaotic reality of the developers. Hence, they are the key knowledge engineers in their respective companies (Nonaka, 1991; Nonaka, 1994; Nonaka & Takeuchi, 1995). These managers were used in this study to answer on behalf of their respective organizations. Thus the unit of analysis, from which original data observations were obtained, was the software organization. Within this study, *a software organization is defined as a whole company or an independent business unit inside a larger company that has software development as its primary business*.

The objective was to develop an instrument that could be used in companies of different sizes, developing either software or combined software and hardware products. Therefore, managers were chosen from companies with corporate membership either in the Association of the Norwegian Software Industry (PROFF) or the Norwegian IT Technology Forum (ITUF). Taken together, these two organizations were considered to be representative for software development within the Norwegian IT industry.

Sample size is an important consideration in the discussion of the internal consistency of measurement scales. Thus, for the purpose of constructing a measurement instrument with satisfactory psychometric properties, the item analysis is the dimensioning factor. Generally, the item analysis to choose a set of items that form an internally consistent scale, requires a sample size of about 100 to 200 respondents (Spector, 1992). Based on these guidelines, our target sample size for ensuring adequate item analysis was a *minimum of 100 respondents*.

A random sample of 154 software and quality managers from the membership lists of the two associations were contacted by telephone to request participation in the study prior to mailing the questionnaires. All managers agreed to participate in the study. We provided the respondents with self-addressed, stamped return envelopes. Also, by keeping the questionnaire as short as possible (the pilot study showed that respondents needed about 10 minutes to complete it), we combined several well-proven techniques for improving the response rate of mailed questionnaires (e.g. Kanuk and Berenson, 1975; Fink and Kosecoff, 1998; Neuman, 2000).

A total of 120 software and quality managers representing whole organizations or independent business units within 55 software companies completed and returned the questionnaire. This represents an effective response rate of 77.9 percent, which is well within the norm of $60+/-20$ for representatives of organizations and mid-level managers suggested for academic studies by Baruch (1999), and considerably higher than prior mail surveys conducted in the Norwegian software and IT industry. As a comparison, a recent study by Larsen and Kautz (1997) on the status of quality assurance and software process improvement in Norway obtained a response rate of 13.3 percent, while Stålhane, Borgersen and Arnesen's (1997) study of the customer's view of software product quality achieved a response rate of 8.4 percent. Given the high response rate in this study, no further analysis was done on the differences between respondents and non-respondents.

As shown in Table 9, the respondent's companies represent a wide variety of industry sectors. However, the sample is, as expected, biased in favor of IT companies (56.7% of the sample). To a large extent, the organizations develop software for external customers (approximately 96% of the sample). Furthermore, the sample shows a mix of both small and

Table 9. Characteristics of the survey sample.

Characteristic		Frequency	Percent
I.	*Characteristics of the Respondent*		
	Average number of years worked in the company	8.4	
	Average number of years worked with software development	11.4	
	Highest completed education		
	Bachelor's degree	38	31.7%
	Master's degree	74	61.7%
	Doctoral degree	5	4.2%
	Other	3	2.5%
	Job function		
	Software manager	95	79.2%
	Quality manager	25	20.8%
II.	*Characteristics of the Respondent's Company*		
	Number of software developers		
	Less than or equal to 30	45	37.5%
	Between 30 and 200	31	25.8%
	More than or equal to 200	44	36.7%
	Primary industry group		
	Public sector	7	5.8%
	Banking/finance/insurance	12	10.0%
	Manufacturing	21	17.5%
	IT sector	68	56.7%
	Other	12	10.0%
	Type of product business		
	Standard applications (shelfware)	31	25.8%
	Tailor made solutions for external customers	84	70.0%
	Tailor made solutions for internal company customers	5	4.2%
	Quality system in use		
	Yes	86	71.7%
	No	34	28.3%

large organizations, with approximately one third (37.5%) of the organizations having 30 or less developers and approximately one third (36.7%) having 200 or more. A large majority of the respondent's organizations have a quality system in use (71.7%). The average length of the respondents' job tenure at their current organization is 8.4 years, while professional tenure (years in software development) is 11.4 years. Two thirds of the respondents (65.9%) holds a master's or doctoral degree.

In the next two sections, we describe the reliability analysis performed to refine the measurement items of the facilitating factors of SPI success.

7. Reliability of Measurement Scales

Reliability refers to the consistency and stability of a score from a measurement scale (Anastasi and Urbina, 1997). Since all types of reliability are concerned with the degree of consistency or agreement between two independently derived sets of scores, they can all be expressed in terms of a *correlation coefficient*. Essentially, the correlation coefficient (r) indicates the degree of correspondence, or relationship, between two sets of scores. The

Table 10. Classification of reliability estimation methods (adapted from Anastasi and Urbina, 1997).

	Scale Forms Required	
Administrations Required	**One**	**Two**
One	Split-Halves Internal Consistency (Coefficient Alpha)	Alternate-Form (immediate)
Two	Test-Retest	Alternate-Form (delayed)

absolute value of the correlation coefficient can range from 0 to 1.0, with 1.0 perfectly reliable and 0 perfectly unreliable (Anastasi and Urbina, 1997).

The types of reliability computed in practice are relatively few. They are summarized in Table 10 in terms of the different methods for measuring reliability, seen in relation to the number of scale forms required and to the number of scale administrations required. Except for the internal consistency method, the other methods in Table 10 have major limitations for field studies like the one presented in this paper. The test-retest method requires two independent administrations of the same measurement scales on the same group of people. The alternate form method requires the administration of two equivalent scales to the same individuals, with or without a time interval. The split-halves method works well in field studies because it requires only a single administration of a single form of a scale. However, its problem is how to split the test to obtain the most nearly equivalent halves, because the estimate of the reliability coefficient totally depends on how the items are split.

An internal consistency technique that overcomes the shortcomings of the split-half method is known as *coefficient alpha* (Cronbach, 1951). This technique computes the mean reliability coefficient estimates for all possible ways of splitting a set of items in two. Hence, coefficient alpha expresses the degree to which items in a scale are homogeneous. The formula for coefficient alpha is given as follows (Cronbach, 1951):

$$\alpha = \left(\frac{n}{n-1}\right)\frac{SD_y^2 - \sum(SD_i^2)}{SD_y^2}$$

In this formula, α is the reliability coefficient for the whole scale, n is the number of items in the scale, SD_y is the standard deviation of total scores on the scale, and SD_i is the standard deviation of individual item scores. Coefficient alpha varies between 0 and 1.0. If there is no true score, but only error in the items, then the sum of variances of the individual items will be the same as the variance of the sum and, consequently, alpha will be equal to zero. If, on the other hand, all items are perfectly reliable and measure the same concept, then coefficient alpha will be equal to one. Coefficient alpha is a conservative estimate that can be regarded as a *lower bound* of reliability (Novick and Lewis, 1967; Carmines and Zeller, 1979).

An internal consistency analysis was performed for each of the six key facilitating factors of SPI success using the SPSS reliability program (SPSS, 1999a, 1999b). We analyzed the

Table 11. Reliability analysis results for the key factors of SPI success.

Key Factors of SPI Success	Item Numbers	Number of Items	Items Deleted	Alpha
Business Orientation	1–5	5	None	.81
Leadership Involvement	6–10	5	None	.87
Employee Participation	11–17	7	None	.80
Concern for Measurement	18–23	6	None	.81
Exploitation of existing knowledge	24–29	6	No. 29	.78
Exploration of new knowledge	30–37	8	None	.85

correlation of items within each scale (item-item), the corrected item-to-total (item-scale) correlations, the effects of reliability if the item was deleted, and the item standard deviation scores to determine which items were candidates for deletion from the scale. Candidate items for deletion were items with low item-item and item-scale correlations, which would raise alpha if deleted. Other candidate items were items with low variance, which would have low explanatory power. However, before any item was deleted, we made a check to ensure that the domain coverage of the construct would not suffer.

This analysis revealed that to obtain satisfactory values for coefficient alpha while retaining the domain coverage only required one item to be eliminated from the exploitation scale (marked with an asterisk in Table 7). Table 11 reports the original sets of measurement items associated with the key factors, the items dropped from the original sets to increase alpha, and the reliability coefficients for the resulting scales.

Table 11 shows that the reliability coefficients ranged from 0.78 to 0.87, indicating that some scales are more reliable than others are. A satisfactory level of reliability will depend on how a measure is being used. However, in the early stages of instrument research, reliabilities of 0.7 or higher are considered sufficient for narrow constructs (Cronbach, 1951; Nunnally and Bernstein, 1994), and 0.55 to 0.70 for moderately broad constructs (Van de Ven and Ferry, 1980). Furthermore, reliability coefficients of 0.50 or above are considered acceptable to compare groups (Fink and Kosecoff, 1998). Besides, Nunnally and Bernstein (1994) argued that increasing reliabilities beyond 0.80 in basic research is often wasteful. Based on these recommendations, our target level of minimum reliability was set in the 0.70 to 0.80 range. Hence, all scales developed for the facilitating factors in this study were judged to be reliable. In the next section, we describe the detailed item analysis performed to evaluate the appropriateness of each item in each scale.

8. Detailed Item Analysis

We made a detailed item analysis based on Nunnally's method to evaluate the assignment of items to scales (Nunnally, 1978; Nunnally and Bernstein, 1994). The method considers the correlation of each item with each scale, in order to determine whether an item belongs to the scale as assigned, whether it belongs to some other scale, or whether it should be eliminated. Compared to items with relatively low correlation with total scores, those that have higher correlation with total scores have more variance relating to the common factor

among items, and they add more to instrument reliability (Nunnally, 1978). Hence, an item that does not correlate highly with any of the scales, should be eliminated.

There are two types of item-scale correlation. The *corrected item-scale correlation* correlates the item being evaluated with all the scale items, excluding itself, while the *uncorrected item-scale correlation* correlates the item in question with the entire set of candidate items, including itself. Although the uncorrected item-scale correlation makes good conceptual sense, the item's inclusion in the scale can inflate the correlation coefficient, thus making it spuriously high. The fewer the number of items in the scale, the bigger the difference that the inclusion or exclusion of the item under examination will make. Therefore, we followed the general advice of examining the corrected item-scale correlations rather than the uncorrected (DeVellis, 1991; Nunnally and Bernstein, 1994).

After eliminating the items in the previous steps, the remaining items were correlated with the total scores of each scale, excluding itself. The corrected item-scale correlation matrix for the facilitating factors of SPI success in Table 12 shows that all items have high correlations with the scales to which they were originally assigned. Furthermore, all correlations are above the cutoff of 0.3 recommended by Nunnally and Bernstein (1994).

With the exception of the "feedback" item (item no. 23), all item-scale correlations are substantially higher than the corresponding correlations with all other scales. However, based on the literature review, the expert review and our own experience, we decided to keep the feedback item within the measurement scale as originally assigned. There are two reasons for this. First, feedback is generally considered as being of utmost importance in all SPI work, and second, we regard feedback as conceptually closer to "Concern for Measurement" rather than to "Exploitation of Existing Knowledge." Therefore, we concluded that all items had been assigned to the appropriate scale, and that no additional items should be deleted.

9. Validity of Measurement Scales

For a scale to be valid, it must also be reliable. Validity is differentiated from reliability in that the former relates to accuracy, while the latter relates to consistency. A measurement scale is valid if it does what it is supposed to do and measures what it is supposed to measure (Cronbach, 1971; Anastasi and Urbina, 1997). If a scale is not valid, it is of little use because it is not measuring or doing what it is supposed to be doing. Three kinds of validity are of special concern for this study. These validity concerns are outlined in Table 13 and discussed next.

9.1. Content Validity

Content validity has to do with the degree to which the scale items represent the domain of the concept under study. Essentially, procedures for content validation involve the systematic examination of the instrument content to determine if it covers a representative sample of the behavior domain to be measured (Davis, 1996). Content validity is built into a test from the outset through the choice of appropriate items (Anastasi and Urbina, 1997).

Table 12. Corrected item-scale correlation matrix.

Key Factors of SPI Success	Item No.	Scale 1	2	3	4	5	6
Business Orientation	1	*.61*	.52	.47	40	.33	−.03
(Scale 1)	2	*.67*	.43	.37	.36	.47	.18
	3	*.49*	.40	.39	.29	.47	.18
	4	*.69*	.53	.24	.28	.51	.21
	5	*.56*	.51	.46	.40	.38	.15
Leadership Involvement	6	.45	*.67*	.24	.33	.42	.09
(Scale 2)	7	.54	*.81*	.37	.41	.39	.05
	8	.49	*.69*	.30	.33	.34	.13
	9	.55	*.70*	.29	.35	.45	.18
	10	.55	*.65*	.30	.25	.30	.00
Employee Participation	11	.09	.18	*.47*	−.01	.20	.35
(Scale 3)	12	.15	.18	*.55*	.10	.28	.41
	13	.28	.28	*.58*	.02	.39	.31
	14	.09	.12	*.47*	.07	.17	.30
	15	.52	.40	*.54*	.39	.45	.25
	16	.48	.34	*.55*	.34	.44	.32
	17	.16	.19	*.53*	.15	.24	.22
Concern for Measurement	18	.32	.32	.02	*.40*	.13	−.06
(Scale 4)	19	.33	.26	.06	*.61*	.25	−.06
	20	.33	.28	.16	*.72*	.36	.20
	21	.33	.33	.17	*.77*	.29	.18
	22	.43	.36	.34	*.61*	.39	.13
	23	.19	.19	.26	*.31*	.31	.16
Exploitation of existing knowledge	24	.34	.28	.35	.16	*.55*	.20
(Scale 5)	25	.52	.38	.43	.34	*.69*	.15
	26	.32	.33	.22	.28	*.52*	−.05
	27	.45	.32	.28	.51	*.54*	.10
	28	.44	.39	.41	.20	*.47*	.19
Exploration of new knowledge	30	.26	.26	.27	.09	.20	*.59*
(Scale 6)	31	.13	.08	.41	.14	.17	*.67*
	32	.15	.09	.30	.08	.16	*.66*
	33	.10	.05	.36	−.03	.13	*.61*
	34	.14	.14	.44	.30	.20	*.53*
	35	−.07	−.16	.24	−.06	−.13	*.57*
	36	.02	−.03	.15	−.07	−.11	*.50*
	37	.25	.20	.38	.28	.26	*.54*

Our procedure for instrument development followed the general recommendations of Cronbach (1971) and Straub (1989) and included:

1. An exhaustive search of the literature for all possible items to be included in the scales.
2. An exploratory study in representative companies to find possible items and scales.
3. Review of the proposed scales by experts of both psychological testing and SPI. There were also asked for suggestions as to any additions or deletions to the scales.

Table 13. Three basic types of validity in measurement instruments.

Types of Validity	Definitions
Content validity	The degree to which the items in the measurement instrument represent the domain or universe of the processes under study.
Construct validity	The degree to which the measurement instrument represents and acts like the processes being measured.
Criterion validity	The degree to which the measurement instrument is able to predict a variable that is designated a criterion.

4. Pilot test of the scales on a set of respondents similar to the target population. These respondents were also encouraged to include suggestions and criticisms to the contents and wording of the scales.

Hence, we argue that our six measurement scales representing the facilitating factors of SPI success developed in this study have content validity since selection of measurement items was based on generally accepted procedures to obtain content validation.

9.2. Construct Validity

Construct validity is an operational concept that examines whether the measurement scales represent and act like the attributes being measured (Cronbach, 1971; Nunnally and Bernstein, 1994). Assuming that the total score of a scale is valid, the extent to which an individual item measures the same thing as the total score is an indicator of the validity of that item. Furthermore, factor analysis is considered to be "a powerful and indispensable method of construct validation" (Kerlinger, 1986). Thus, in addition to item-total correlations, the construct validity of the scales in the measurement instrument was assessed using factor analysis.

Although many investigators advocate the construction of measurement instruments through factor analysis, Spector (1992) and Nunnally and Bernstein (1994) argued that factor analysis should rather be used *after* the instrument is constructed. The main reason for this is that the construction of a measurement instrument should be guided by theories, rather than by random efforts to relate things to one another through "shotgun statistics." This fits the approach in this research where the constructs of interest are based on a substantial body of prior research and have been explicated prior to any item development.

The construct validity of the six measurement scales was evaluated by analyzing the items of each of the scales using principal components analysis. However, this analysis, as all factor-analysis procedures, includes subjective judgement when it comes to determining the number of factors and their interpretation (Spector, 1992).

Several guidelines have been developed to assist in deciding how many factors to extract. The two most widely used criteria are Kaiser's (1960, 1970) *eigenvalue rule* and Cattell's (1966) *scree test*. The eigenvalue rule is based on retaining only factors that explain more variance than the average amount explained by one of the original items (i.e. components

Table 14. Summary of factor matrices for each construct.

Scale	Eigenvalue	Item Loading Range for Component 1	# Items with Loadings > 0.6
1. Business orientation	2.9	.72 to .82	5 (out of 5)
2. Leadership involvement	3.3	.78 to .89	5 (out of 5)
3. Participation	3.2	.62 to .72	7 (out of 7)
4. Measurement	3.1	.43 to .83	4 (out of 6)
5. Exploitation	2.7	.65 to .84	5 (out of 5)
6. Exploration	3.9	.61 to .78	8 (out of 8)

with eigenvalue > 1). However, when Zwick and Velicer (1986) examined the effect of different criteria for extracting factors, they found that Kaiser's method tended to severely overestimate the number of components. Furthermore, Cliff (1988) called the rational behind the eigenvalue rule into question, and despite its wide use, Nunnally and Bernstein (1994) did not recommend it because it tends to suggest too many factors. The scree test is based on a subjective examination of the plot of eigenvalues for each successive factor, looking for an "elbow" in the plot. Cattell's guidelines call for retaining those factors above the "elbow" and rejecting those below it.

Each item also has a final loading on each factor. These loadings are the correlations between items and the factors, thus the square of a loading can be seen as the amount of variance common to both. Comrey (1973) suggested that loadings in excess of 0.45 could be considered fair, those greater than 0.55 as good, those of 0.63 very good, and those of 0.71 as excellent. For a study with a relatively small number of cases, the quality of a factor must be assessed both in terms of the number and size of its loadings. Stevens (1992) suggested that a reliable factor must have four or more loadings of at least 0.6 when the number of cases is below 150.

We used a combination of eigenvalues, cut-off points of the scree plots and factor loadings as a guide for interpreting the dimensionality of the scales. Furthermore, each scale was assumed to be a separate construct. Table 14 shows the eigenvalues and item loading ranges for each scale. Analysis of the eigenvalues showed that five of the six scales formed a single factor. In the case of the exploration scale, two components seemed to emerge with eigenvalues greater than 1.0. However, the eigenvalue of the second factor was only slightly above this threshold (1.02).

Next, we examined both the scree plot and the item loadings in order to further interpret the dimensionality of the exploration scale. The scree plot showed a clear break after the first component (see Figure 2). Furthermore, all item loadings for this component were greater than 0.6. The combined result of these analyses indicates fairly strong support for the hypothesis that the exploration scale can be considered as one construct.

In brief, the results of the factor analysis support the view that each of the measurement scales in the questionnaire has a high degree of unidimensionality. Hence, there is tentative evidence of construct validity for all six scales.

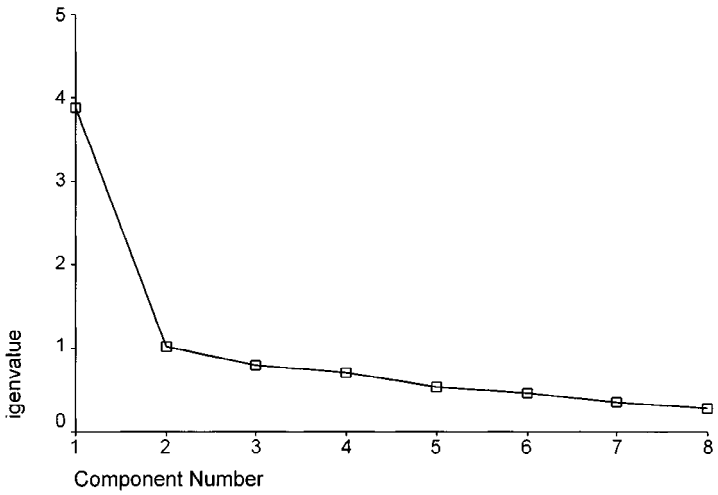

Figure 2. Scree plot for the Exploration scale.

9.3. Criterion Validity

Criterion-related validity, sometimes called external validity, is concerned with the degree to which the scales under study are related to an independent measure of the relevant criterion (Anastasi and Urbina, 1997).

Organizational performance is the ultimate criterion for any SPI activity. The six facilitating factors, thus, have criterion-related validity if, taken collectively, they are highly and positively correlated with performance in a software organization. Performance is a complex construct, however, reflecting the criteria and standards used by decision-makers to assess the functioning of a software organization. That is, performance is a value judgement on the results desired from an organization (Van de Ven and Ferry, 1980).

Investments in SPI share many features with research and development investments, frequently having corporate-wide, intangible and long lasting effects. Therefore, quantitative measures and financial estimates tend to be difficult to obtain and easy to manipulate. The main distinction between the objectivity of different success measures is not a matter of using quantitative or financial figures as opposed to developers' or managers' perceptions. For example, the return on investment, net present value and payback periods are often regarded as objective measures. However, because of the many difficulties in predicting and assessing costs, and especially benefits, such investment analyses are usually based on experts' subjective judgement (Saarinen 1996; Zahran, 1998). In many cases, the subjective measures may be better than the financial measures (Saarinen, 1996). This is mainly due to complicated contingencies that are best judged by those who are responsible for the object of evaluation.

Furthermore, value judgements on desirable performance outcomes often change over time in a way that threatens the applied relevance of longitudinal assessments of orga-

nizations. Finally, decision makers often disagree on a given set of criteria to evaluate organizational performance. However, performance measurement does not require a consensus on effectiveness goals, criteria, and standards. An organizational assessment simply requires that the definitions of performance be made explicit and that the researcher determines whose value judgements and criteria will be operationalized and measured (Van de Ven and Ferry, 1980).

Based on the preceding discussion, and that subjective performance measures are widely used and accepted in organizational research (Lawrence and Lorsch, 1986; Powell, 1995), we considered a subjective evaluation approach with multi-item constructs as an appropriate way to assess organizational performance.

Generally, the intent of SPI is increased competitive advantage through improved software quality, increased productivity, and decreased lead-time for product development. In other words, SPI should lead to "better, faster, [and] cheaper software development" (Sanders, 1998). Based on these general recommendations, we operationalized and measured SPI success along two dimensions, based on two multi-item measures. Thus, each manager was asked to rate, on 5-point bipolar Likert scales, (1) the level of perceived SPI success and (2) the performance of their organization for the past three years with respect to cost reduction, cycle time reduction, and customer satisfaction. This is similar to the approach used by Teng, Jeong and Grover (1998) to measure reengineering project success and the relationship between characteristics of reengineering projects and implementation success.

Two items were used to measure the level of perceived SPI success, while three items were used to measure the level of organizational performance. As for the facilitating factors, all items were written specifically for this study. These items are included in the validated measurement instrument in Appendix B. The ratings for the two performance dimensions were averaged to form a single measure of overall SPI success. The reliability coefficient for the five items of the SPI success measure was 0.76.

The criterion validity of the measurement instrument was found by computing the multiple correlation (R) between SPI success and the measures of the six facilitating factors. The multiple correlation coefficient was 0.76 and the adjusted R-square was 0.56, thus, explaining 56 percent of the variation in SPI success. Furthermore, the F-value of 25.95 was highly significant ($p < 0.0005$).

Cohen (1988) defined the effect size index, f^2, for the squared multiple correlation coefficient, R^2, where $f^2 = R^2/(1-R^2)$. Also, he defined a large effect size as $f^2 \geq 0.35$. Given the relationship between the effect size index and the squared multiple correlation, a large effect size of $f^2 = 0.35$ corresponds to a squared multiple correlation of $R^2 = 0.26$ and a multiple correlation coefficient of $R = 0.51$. In other words, the results indicate that *a large amount of the variance in SPI success has been highly significantly explained by the six facilitating factors.* Thus, taken together, the six factors have a high degree of criterion-related validity.

10. Conclusions

In this paper we have described an extensive investigation for the development of an instrument for measuring the key factors of success in SPI. The major premise was that it is

critical to understand the important factors affecting SPI success in order to improve software processes. However, central to this understanding is the development of an instrument for measuring these factors.

The results of reliability and validity analyses show that the instrument has desirable psychometric properties. However, demonstrating validity is a continuous process, and validity is a matter of degree rather than an all-or-none property. Moreover, one validates the *use* to which a measurement instrument is put rather than the instrument itself (Nunnally and Bernstein, 1994). Consequently, it is only through the accumulation of data from multiple studies of instrument usage that we can make strong claims of validity.

Acknowledgments

This work was supported in part by the Research Council of Norway under Grant No. 118206/221 and by SINTEF Telecom and Informatics. The author would like to thank Reidar Conradi, Tor Stålhane, Stan Rifkin and the anonymous reviewers for their valuable comments on earlier drafts of this paper.

References

Ahire, S. L., Golhar, D. Y. and Waller, M. A. 1996 Development and Validation of TQM Implementation Constructs, *Decision Sciences*, 27(1), 23–56.

Anastasi, A. and Urbina, S. 1997 *Psychological Testing*, Seventh edition, Upper Saddle River, New Jersey: Prentice-Hall.

Argyris, C. and Schön, D. A. 1978 *Organizational Learning: A Theory of Action Perspective*, Reading, Massachusetts: Addison-Wesley.

Argyris, C. and Schön, D. A. 1996 *Organizational Learning II: Theory, Method, and Practice*, Reading, Massachusetts: Addison-Wesley.

Bagozzi, R. P. 1996 Measurement in Marketing Research: Basic Principles of Questionnaire Design, in R. P. Bagozzi (Ed.), *Principles of Marketing Research*, Cambridge, Massachusetts: Blackwell, pp. 1–49.

Baruch, Y. 1999 Response Rate in Academic Studies—A Comparative Analysis, *Human Relations*, 52(4), 421–438.

Basili, V. R. 1996 The Role of Experimentation in Software Engineering: Past, Current, and Future, *Proceedings of the 18th International Conference on Software Engineering (ICSE-18)*, Berlin, Germany, March 25–29, pp. 442–449.

Basili, V. R. and Caldiera, G. 1995 Improve Software Quality by Reusing Knowledge and Experience, *Sloan Management Review*, 37, Autumn, pp. 55–64.

Basili, V. R. and Rombach, H. D. 1988 The TAME Project: Towards Improvement-Oriented Software Environments, *IEEE Transactions on Software Engineering*, 14(6), 758–773.

Basili, V. R. and Selby, R. W. 1991 Paradigms for Experimentation and Empirical Studies in Software Engineering, *Reliability Engineering and System Safety*, 32(1–2), pp. 171–191.

Basili, V. R., Selby, R. W. and Hutchens, D. H. 1986 Experimentation in Software Engineering, *IEEE Transactions on Software Engineering*, 12(7), 733–743.

Black, S. A. and Porter, L. J. 1996 Identification of the Critical Factors of TQM, *Decision Sciences*, 27(1), 1–21.

Blalock, H. M. 1969 *Theory Construction: From Verbal to Mathematical Formulations*, Englewood Cliffs, N.J.: Prentice-Hall.

Carmines, E. G. and Zeller, R. A. 1979 *Reliability and Validity Assessment*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-017, Newbury Park, CA: Sage.

Cattell, R. B. 1966 The Scree Test for the Number of Factors, *Multivariate Behavioral Research*, 1, 245–276.

Choo, C. W. 1995 The Knowing Organization: How Organizations Use Information to Construct Meaning Create Knowledge and Make Decisions, *International Journal of Information Management*, 16(5), 329–340.
Choo, C. W. 1998 *The Knowing Organization: How Organizations Use Information to Construct Meaning, Create Knowledge, and Make Decisions*, New York: Oxford University Press.
Cliff, N. R. 1988 The Eigenvalues-greater-than-one Rule and the Reliability of Components, *Psychological Bulletin*, 103, 276–279.
Cohen, J. 1988 *Statistical Power Analysis for the Behavioral Sciences*, Second Edition, Hillsdale, New Jersey: Laurence Erlbaum.
Comrey, A. 1973 *A First Course on Factor Analysis*, London: Academic Press.
Cronbach, L. J. 1951 Coefficient Alpha and the Internal Consistency of Tests, *Psychometrica*, 16, 297–334, September.
Cronbach, L. J. 1971 Test Validation, in *Educational Measurement*, 2nd Edition, R. L. Thorndike (ed.), American Council on Education, Washington, D.C, pp. 443–507.
Crosby, P. B. 1979 *Quality is Free: The Art of Making Quality Certain*, New York: McGraw-Hill.
Crosby, P. B. 1984 *Quality Without Tears*, New York: McGraw-Hill.
Crosby, P. B. 1996 *Quality Is Still Free: Making Quality Certain in Uncertain Times*, New York: McGraw-Hill.
Davis, D. 1996 *Business Research for Decision Making*, Fourth Edition, Belmont, California: Duxbury Press.
Deming, W. E. 1982 *Quality, Productivity, and Competitive Position*, Camebridge, Massachusetts: MIT Centre for Advanced Engineering Study.
Deming, W. E. 1986 *Out of the Crisis*, Cambridge, Massachusetts: MIT Center for Advanced Engineering Study.
Deephouse, C., Mukhopadhyay, T., Goldenson, D. R. and Kellner, M. I. 1996 Software Processes and Project Performance, *Journal of Management Information Systems*, 12(3), 187–205.
DeVellis, R. F. 1991 *Scale Development: Theory and Applications*, Newbury Park, CA: Sage.
Dybå, T. 2000a Improvisation in Small Software Organizations, *IEEE Software*, 17(5), Sept.-Oct.
Dybå, T. (Ed.) 2000b *SPIQ—Software Process Improvement for better Quality: Methodology Handbook*, Version 3.0 (Final), IDI Report 2/2000, Norwegian University of Science and Technology, Trondheim, Norway (in Norwegian).
EFQM 1999 *The EFQM Excellence Model*, Brussels: European Foundation for Quality Management.
El Emam 1998 The Internal Consistency of the ISO/IEC 15504 Software Process Capability Scale, *Proceedings of the Fifth International Symposium on Software Metrics*, IEEE Computer Society Press, pp. 72–81.
El Emam, K. and Birk, A. 2000 Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability, *IEEE Transactions on Software Engineering*, 26(6), 541–566.
El Emam, K. and Madhavji, N. H. 1995 The Reliability of Measuring Organizational Maturity, *Software Process—Improvement and Practice*, 1, 3–25.
El Emam, K. and Madhavji, N. H. 1996 An Instrument for Measuring the Success of the Requirements Engineering Process in Information Systems Development, *Empirical Software Engineering*, 1(3), 201–240.
El Emam, K., Fusaro, P. and Smith, B. 1999 Success Factors and Barriers for Software Process Improvement, in R. Messnarz, and C. Tully (Eds.) *Better Software Practice for Business Benefit: Principles and Experiences*, Los Alamitos, California: IEEE Computer Society Press, pp. 355–371.
El Emam, K., Goldenson, D. R., McCurley J. and Herbsleb, J. 1998 *Success or Failure? Modeling the Likelihood of Software Process Improvement*, Technical Report, ISERN-98-15, International Software Engineering Research Network.
Feigenbaum, A. V. 1991 *Total Quality Control*, Fortieth Anniversary Edition, New York: McGraw-Hill.
Fink, A. and Kosecoff, J. 1998 *How to Conduct Surveys: A Step-By-Step Guide*, Second Edition, Thousand Oaks, California: Sage Publications.
French, W. L. and Bell, C. H. Jr. 1999 *Organization Development: Behavioral Science Interventions for Organization Improvement*, Sixth Edition, Upper Saddle River, New Jersey: Prentice-Hall.
Fusaro, P., El Emam, K. and Smith, B. 1998 The Internal Consistency of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension, *Empirical Software Engineering*, 3(2), 179–201.
Garvin, D. A. 1983 Quality on the Line, *Harvard Business Review*, 61(5), 65–75.
Garvin, D. A. 1984 Japanese Quality Management, *Columbia Journal of World Business*, 19(3), 3–19.
Garvin, D. A. 1993 Building a Learning Organization, *Harvard Business Review*, 71(4), 78–91.
Goldenson, D. R. and Herbsleb, J. (1995) *After the Appraisal: A Systematic Survey of Process Improvement, its Benefits, and Factors that Influence Success*, Technical Report, CMU/SEI-95-TR-009, Carnegie Mellon University, Software Engineering Institute.
Goldenson, D. R., El Emam, K., Herbsleb, J. and Deephouse, C. 1999 Empirical Studies of Software Process

Assessment Methods, in K. El Emam and N. H. Madhavji (Eds.), *Elements of Software Process Assessment and Improvement*, Los Alamitos, California: IEEE Computer Society Press, pp. 177–218.
Guilford, J. P. 1954 *Psychometric Methods*, 2nd edition, New York: McGraw-Hill.
Harrison, R., Badoo, N., Barry, E., Biffl, S., Parra, A., Winter, B. and Wuest, J. 1999 Directions and Methodologies for Empirical Software Engineering Research, *Empirical Software Engineering*, 4(4), 405–410.
Humphrey, W. S. 1989 *Managing the Software Process*, Reading, Massachusetts: Addison-Wesley.
Humphrey, W. S. 1997 *Managing Technical People: Innovation, Teamwork, and the Software Process*, Reading, Massachusetts: Addison-Wesley.
Hunt, R. and Buzan, T. 1999 *Creating a Thinking Organization: Groundrules for Success*, Hampshire, England: Gower.
Hunter, J. E. and Schmidt, F. L. 1990 *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*, Newbury Park, California: Sage Publications.
Ishikawa, K. 1986 *Guide to Quality Control*, Second Edition, New York: Quality Resources.
Ishikawa, K. 1990 *Introduction to Quality Control*, London: Chapman & Hall.
ISO/DIS 9000 2000 *Quality Management Systems—Fundamentals and Vocabulary*.
ISO/DIS 9004 2000 *Quality Management Systems—Guidelines for Performance Improvements*.
ISO/IEC TR 15504-7 1998 *Information Technology—Software Process Assessment—Part 7: Guide for Use in Process Improvement*.
Jeffery, D. R. and Votta, L. G. 1999 Guest Editor's Special Section Introduction, *IEEE Transactions on Software Engineering*, 25(4), 435–437.
Juran, J. M. 1992 *Juran on Quality by Design: The New Steps for Planning Quality into Goods and Services*, New York: Free Press.
Juran, J. M. and Godfrey, A. B. (Eds.) 1999 *Juran's Quality Handbook*, Fifth Edition, New York: McGraw-Hill.
Kaiser, H. F. 1960 The Application of Electronic Computers to Factor Analysis, *Educational and Psychological Measurements*, 20, 141–151.
Kaiser, H. F. 1970 A Second Generation Little Jiffy, *Psychometrika*, 35, 401–417.
Kanuk, L. and Berenson, C. 1975 Mail Surveys and Response Rates: A Literature Review, *Journal of Marketing Research*, 12, 440–453.
Kerlinger, F. 1986 *Foundations of Behavioral Research*, Holt, Rinehart and Winston.
Larsen, E. Å, and Kautz, K. 1997 Quality Assurance and Software Process Improvement in Norway, *Software Process—Improvement and Practice*, 3(2), 71–86.
Lawrence, P. R. and Lorsch, J. W. 1986 *Organization and Environment: Managing Differentiation and Integration*, 2nd Edition, Boston: Harvard Business School Press.
Likert, R. 1932 A Technique for the Measurement of Attitudes, *Archives of Psychology*, 22(140).
Likert, R. 1967 *The Human Organization: Its Management and Value*, New York: McGraw-Hill.
Likert, R. and Roslow, S. 1934 *The Effects Upon the Reliability of Attitude Scales of Using Three, Five or Seven Alternatives*, Working Paper, New York University.
Lissitz, R. W. and Green, S. B. 1975 Effects of the Number of Scale Points on Reliability: A Monte Carlo Approach, *Journal of Applied Psychology*, 60, 10–13, February.
March, J. G. 1991 Exploration and Exploitation in Organizational Learning, *Organization Science*, 2(1), 71–87.
March, J. G. 1999 *The Pursuit of Organizational Intelligence*, Malden, Massachusetts: Blackwell.
Miller, G. A. 1956 The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information, *Psychological Review*, 63, 81–97.
Neuman, W. L. 2000 *Social Research Methods: Qualitative and Quantitative Approaches*, Fourth Edition, Boston: Allyn and Bacon.
Nevis, E. C., DiBella, A. J. and Gould, J. M. 1995 Understanding Organizations as Learning Systems, *Sloan Management Review*, 37, 73–85.
Nonaka, I. 1991 The Knowledge-Creating Company, *Harvard Business Review*, 69(6), 96–104.
Nonaka, I. 1994 A Dynamic Theory of Organizational Knowledge Creation, *Organization Science*, 5(1), 14–37.
Nonaka, I. and Takeuchi, H. 1995. *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*, New York: Oxford University Press.
Novick, M. and Lewis, G. 1967 Coefficient Alpha and the Reliability of Composite Measurements, *Psychometrika*, 32, 1–13.
Nunnally, J. C. 1978 *Psychometric Theory*, Second edition, New York: McGraw-Hill.
Nunnally, J. C. and Bernstein, I. A. 1994 *Psychometric Theory*, Third edition, New York: McGraw-Hill.

Powell, T. C. 1995 Total Quality Management as Competitive Advantage: A Review and Empirical Study, *Strategic Management Journal*, 16, 15–37.
Rogers, E. M. 1995 *Diffusion of Innovations*, Fourth Edition, New York: The Free Press.
Sanders, M. (Ed.) 1998 *The SPIRE Handbook: Better, Faster, Cheaper Software Development in Small Organisations*, Dublin: Centre for Software Engineering Ltd.
Saraph, J. V., Benson, P. G. and Schroeder, R. G. 1989. An Instrument for Measuring the Critical Factors of Quality Management, *Decision Sciences*, 20(4), 810–829.
Schaffer, R. H. and Thomson, H. A. 1992. Successful Change Programs Begin with Results, *Harvard Business Review*, 70, 80–89.
Senge, P. M. 1990 *The Fifth Discipline: The Art and Practice of the Learning Organization*, New York: Doubleday.
Senge, P. M., Kleiner, A., Roberts, C., Ross, R. and Smith, B. 1994 *The Fifth Discipline Fieldbook: Strategies and Tools for Building a Learning Organization*, New York: Currency/Doubleday.
Senge, P. M., Kleiner, A., Roberts, C., Ross, R., Roth, G. and Smith, B. 1999 *The Dance of Change: The Challenges of Sustaining Momentum in Learning Organizations*, New York: Currency/Doubleday.
Spector, P. 1992 *Summated Rating Scale Construction: An Introduction*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-082, Newbury Park, CA: Sage.
SPSS 1999a *SPSS Base 9.0: User's Guide*, Chicago, IL: SPSS Inc.
SPSS 1999b *SPSS Base 9.0: Applications Guide*, Chicago, IL: SPSS Inc.
Stelzer, D. and Mellis, W. 1998 Success Factors of Organizational Change in Software Process Improvement, *Software Process—Improvement and Practice*, 4(4), 227–250.
Stelzer, D., Mellis, W. and Herzwurm, G. 1996 Software Process Improvement via ISO 9000? Results of Two Surveys among European Software Houses, *Proceedings of the 29th Hawaii International Conference on Systems Sciences*, January 3–6, Wailea, Hawaii, USA.
Stevens, J. 1992 *Applied Multivariate Statistics for the Social Sciences*, London: Lawrence Erlbaum.
Straub, D. W. 1989 Validating Instruments in MIS Research, *MIS Quarterly*, 13(2), 147–169.
Stålhane, T., Borgersen, P. C. and Arnesen, K. 1997 In Search of the Customer's Quality View, *Journal of Systems Software*, 38, 85–93.
Saarinen, T. 1996 An Expanded Instrument for Evaluating Information System Success, *Information & Management*, 31, 103–118.
Taguchi, G. 1986 *Introduction to Quality Engineering: Designing Quality into Products and Processes*, Tokyo: Asian Productivity Organization.
Taguchi, G., Elsayed, E. A. and Hsiang, T. C. 1989 *Quality Engineering in Production Systems*, New York: McGraw-Hill.
Teng, J. T. C., Jeong, S. R. and Grover, V. 1998 Profiling Successful Reengineering Projects, *Communications of the ACM*, 41(6), 96–102.
Van de Ven, A. H. and Ferry, D. L. 1980 *Measuring and Assessing Organizations*, New York: John Wiley & Sons.
Winter, S. G. 1994 Organizing for Continuous Improvement: Evolutionary Theory Meets the Quality Revolution, in J. A. C. Baum and J. Singh (Eds.) *The Evolutionary Dynamics of Organizations*, Oxford University Press.
Yusof, S. M. and Aspinwall, E. 1999 Critical Success Factors for Total Quality Management Implementation in Small and Medium Enterprises, *Total Quality Management*, 10(4&5), 803–809.
Zahran, S. 1998 *Software Process Improvement: Practical Guidelines for Business Success*, Harlow, England: Addison-Wesley.
Zwick, W. R. and Velicer, W. F. 1986 Comparison of Five Rules for Determining the Number of Components to Retain, *Psychological Bulletin*, 99, 432–442.

Appendix A: Original List of Facilitating Factors and Indicators

The items marked by an asterisk (*) were dropped after the expert review. Items marked by two asterisks (**) were eventually deleted in order to improve the reliability of the final instrument.

Key Factors	**Potential Indicators**
Business Orientation	• amount of specific goals for SPI within the company* • extent of goal-orientation in SPI activities • extent of long term planning of SPI activities* • extent to which SPI goals and policy are understood • degree of integrating SPI actions with "real" work • extent to which SPI goals are aligned with business goals • thoroughness in the priority of improvement actions* • degree of balance between short-term and long-term goals
Leadership Involvement	• degree of participation by management in the SPI process • acceptance of responsibility for SPI by management • extent to which top management has objectives for software performance • degree to which top management is evaluated for software performance* • degree to which management considers SPI as a way to increase competitive advantage • amount of review of SPI issues in top management meetings
Employee Participation	• extent of employee involvement in decisions about what should best be done at their own level (co-determination) • extent of employee participation in development planning** • extent of developer participation in the formalization of routines • extent to which employees contribute with improvement proposals • extent of employee interaction with customers and suppliers* • extent to which employees have responsibility for SPI • amount of training in SPI concepts* • extent of developer participation in goal-setting, data-analysis, and interpretation • extent of participatory management* • extent of on-going dialogue and discussion about software development • extent of on-going dialogue and discussion about SPI
Concern for Measurement	• importance of measuring performance • availability of quality data (defects, timeliness, etc) • extent to which quality data are available to developers • extent to which quality data are available to managers • extent to which quality data, charts, etc. are displayed in corridors etc* • extent to which simple statistical tools are used* • amount of feedback provided to developers on their performance

Key Factors	Potential Indicators
Exploitation of existing knowledge	• extent to which existing knowledge is exploited • extent of learning from past experience • degree of elaboration of existing ideas, strategies, and technologies* • degree of systemization of past experience • importance of formulating experience into explicit concepts* • degree to which formal routines are based on past experience • extent to which computerized organizational memory are used* • importance of narrow (specialized) skills* • importance of maximum breakdown of task* • extent of searching for stability* • degree of internal experience transfer • extent of risk aversion**
Experimentation	• extent of exploration of new knowledge* • degree of experimentation with new ideas, strategies, and technologies • importance of exploring opportunities* • extent to which innovation/change is encouraged • degree of learning from other organizations* • extent to which risk taking is encouraged* • importance of multiple broad skills* • degree of adaptability to rapid change, increasing complexity and environmental uncertainty • ability to question underlying values
Operational variety	• importance of matching the variety and complexity of the organization's environment • extent of variety in the use of methods, techniques and tools • extent to which diversity is appreciated* • degree of detail in task specifications (minimum critical specification) • degree of flexibility in task execution • importance of optimum variety in task performance* • importance of optimum grouping of tasks*

Appendix B: An Instrument for Measuring the Key Factors of Success in Software Process Improvement

Participants in the study were asked to mark or circle the best response to each statement on the following 5-point scale: "Strongly disagree," "Disagree," "Neither agree nor disagree," "Agree," and "Strongly agree" for each of the following factors.

Business Orientation

1. We have established unambigous goals for the organization's SPI activities.
2. There is a broad understanding of SPI goals and policy within our organization.

3. Our SPI activities are closely integrated with software development activities.
4. Our SPI goals are closely aligned with the organization's business goals.
5. We have a fine balance between short-term and long-term SPI goals.

Leadership Involvement

6. Management is actively supporting SPI activities.
7. Management accepts responsibility for SPI.
8. Management considers SPI as a way to increase competitive advantage.
9. Management is actively participating in SPI activities.
10. SPI issues are often discussed in top management meetings.

Employee Participation

11. Software developers are involved to a great extent in decisions about the implementation of their own work.
12. Software developers are actively contributing with SPI proposals.
13. Software developers are actively involved in creating routines and procedures for software development.
14. We have an on-going dialogue and discussion about software development.
15. Software developers have responsibility related to the organization's SPI activities.
16. Software developers are actively involved in setting goals for our SPI activities.
17. We have an on-going dialogue and discussion about SPI.

Concern for Measurement

18. We consider it as important to measure organizational performance.
19. We regularly collect quality data (e.g. defects, timeliness) from our projects.
20. Information on quality data is readily available to software developers.
21. Information on quality data is readily available to management.
22. We use quality data as a basis for SPI.
23. Our software projects get regular feedback on their performance.

Exploitation of Existing Knowledge

24. We exploit the existing organizational knowledge to the utmost extent.

25. We are systematically learning from the experience of prior projects.
26. Our routines for software development are based on experience from prior projects.
27. We collect and classify experience from prior projects.
28. We put great emphasis on internal transfer of positive and negative experience.
29. To the extent we can avoid it, we do not take risks by experimenting with new ways of working.*

Exploration of New Knowledge

30. We are very capable at managing uncertainty in the organization's environment.
31. In our organization, we encourage innovation and creativity.
32. We often carry out trials with new software engineering methods and tools.
33. We often conduct experiments with new ways of working with software development.
34. We have the ability to question "established" truths.
35. We are very flexible in the way we carry out our work.
36. We do not specify work processes more than what is absolutely necessary.
37. We make the most of the diversity in the developer's skills and interests to manage the variety and complexity of the organization's environment.

Organizational Performance

1. Our SPI work has substantially increased our software engineering competence.
2. Our SPI work has substantially improved our overall performance.
3. Over the past 3 years, we have greatly reduced the cost of software development.
4. Over the past 3 years, we have greatly reduced the cycle time of software development
5. Over the past 3 years, we have greatly increased our customer's satisfaction.

Note:

* Starred items were removed from the final version of the instrument and should not be used.

Tore Dybå is a research scientist at Department of Computer Science at SINTEF (The Foundation for Scientific and Industrial Research at the Norwegian Institute of Technology). He is also a research fellow in computer science at the Norwegian University of Science and Technology working on a doctoral thesis investigating the key learning processes and factors for success in SPI. He received his M.Sc. in Computer Science from the Norwegian Institute of Technology in 1986, and worked as a consultant both in Norway and in Saudi Arabia before he joined SINTEF in 1994. His current research interests include empirical software engineering, software process improvement, and organizational learning.