SOFTWARE PROCESS IMPROVEMENT AND PRACTICE
Softw. Process Improve. Pract. 2001; **6**: 205–242 (DOI: 10.1002/spip.146)

Findings from Phase 2 of the SPICE Trials

Standards Section

Ho-Won Jung,[1*,†] Robin Hunter,[2] Dennis R. Goldenson[3] and Khaled El-Emam[4]
[1] *Department of Business Administration, Korea University, Anam-dong 5Ka, Sungbuk-gu, Seoul 136-701, Korea*
[2] *Department of Computer Science, University of Strathclyde, Richmond Street, Glasgow G1 1XH, UK*
[3] *Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA*
[4] *National Research Council, Canada Institute for Information Technology, Building M-50, Montreal Road, Ottawa, Ontario, Canada K1A OR6*

The international SPICE (Software Process Improvement and Capability dEtermination) project was set up to support the development of the ISO/IEC 15504 standard for software process assessment (SPA). The project mounted a set of trials to validate the emerging standard against the goals and requirements defined at the start of the SPICE project and to verify the consistency and usability of its component parts. A considerable number of empirical evaluation studies have been conducted during the Phase 2 SPICE Trials based on ISO/IEC PDTR 15504 (between September 1996 and June 1998). Such an exercise is unprecedented in the software engineering standards community and it provides a unique opportunity for empirical validation. The purpose of this paper is to present major parts of the findings of the empirical studies conducted as part of the SPICE Project during Phase 2 of the SPICE Trials. The topics covered in this paper include (i) investigation into reasons for performing SPAs, (ii) evaluation of the internal consistency of the capability dimension, (iii) use of interrater agreement as a measure of the reliability of assessments, (iv) evaluation of the predictive validity of process capability, (v) evaluation of an exemplar model (Part 5), (vi) identification of factors influencing assessor effort, and (vii) empirical comparison between ISO/IEC PDTR 15504 and ISO 9001. Major lessons learned as well as future research directions are summarized on the strengths and weaknesses of ISO/IEC 15505. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: ISO/IEC 15504; SPICE Trials; SPA(Software Process Assessment)

1. INTRODUCTION

The SPICE (Software Process Improvement and Capability dEtermination) Project is an ongoing international standardization project that supports the development of the emerging standard ISO/IEC 15504 for software process assessment (SPA)[1]. Unique among software engineering

*Correspondence to: Ho-Won Jung, Department of Business Administration, Korea University, Anam-dong 5Ka, Sungbuk-gu, Seoul 136-701, Korea
†E-mail: hwjung@mail.korea.ac.kr

Copyright © 2001 John Wiley & Sons, Ltd.

[1]SPA in ISO/IEC 15504 has two principal contexts for its use: capability determination and process improvement. This definition is different from an earlier SEI assessment method 'SPA' that is the predecessor of CBA IPI (CMM Based Appraisal for Internal Process Improvement) (Dunaway and Masters 1996).

standardization efforts, the developers of ISO/IEC 15504 deliberately initiated an international effort to empirically evaluate the standard. This effort is known as the SPICE Trials (El-Emam and Goldenson 1995, Maclennan and Ostrolenk 1995, Smith and El-Emam 1996). A considerable number of empirical evaluation studies have been conducted during Phase 2 of the SPICE Trials (September 1996 to June 1998).

The SPICE Trials were conceived, partially, to address concerns within the software engineering community regarding the lack of an evidence to support software engineering standards; in particular that there is lack of an empirical basis that demonstrates the standards do represent 'good' practices. For example, Pfleeger *et al*. (1994) state 'Standards have codified approaches whose effectiveness has not been rigorously and scientifically demonstrated. Rather, we have too often relied on anecdote, 'gut feeling', the opinions of experts, or even flawed research'. Similar arguments can be found in Fenton and Page (1993) and Fenton *et al*. (1993).

The purpose of this paper is to present some of the more significant empirical studies conducted during Phase 2 of the SPICE Trials. We review the empirical results on SPAs to date, attempt to draw conclusions from existing studies, and distill what we have learned thus far. Data analyses of the Phase 2 SPICE Trials are still underway. The topics that were covered in this paper are just some of the findings so far.

Phase 2 trial assessments were conducted with ISO/IEC PDTR[2] 15504. In this paper we sometimes take short cuts and refer to 'ISO/IEC 15504' or '15504'. However, it should always be understood that we are referring to ISO/IEC PDTR 15504.

Seven of the issues investigated empirically during Phase 2 are covered. Section 2 is an overview of the emerging ISO/IEC 15504 standard and Section 3 describes the data collection method and how multiple imputation was performed to compensate for missing data. Section 4 provides a descriptive summary of the Phase 2 SPICE Trials data. In Section 5, reasons for performing SPA are addressed in the first empirical study. Section 6 is concerned with the reliability of the process capability dimension. The use of interrater agreement as a measure of the reliability of assessments is illustrated in benchmark and a case study in Section 7. Section 8 shows that the ISO/IEC 15504 standard is valid as a predictive measure of process capability for ISO/IEC 15504 processes ENG.2 (Development software requirements), ENG.3 (Development software design), ENG.4 (Implement software design) and ENG.5 (Integrate and test software). Section 9 provides an evaluation of the exemplar model known as ISO/IEC 15504: Part 5. In Section 10, two regression models are developed showing the factors that affect assessor effort. Section 11 provides empirical comparison between ISO/IEC 15504 and ISO 9001. Research limitations are described in Section 12. Finally, in Section 13 major lessons learned as well as future research directions are summarized at point of the strengths and weaknesses of the ISO/IEC 15504.

2. OVERVIEW OF ISO/IEC 15504 AND SPICE TRIALS

2.1. ISO/IEC 15504 and the SPICE Trials

In June 1991, the International Standards group for Software Engineering, ISO/IEC JTC1/SC7, approved a study period to investigate the need and requirements for a standard for software process assessment (Resolution 144). The results of the international study, which are contained in a study report (ISO/IEC N944R 1992), present the following major conclusions:

- There is international consensus on the need and requirements for a standard for process assessment.
- There is international consensus on the need for a rapid route to development and trialling to provide usable output in an acceptable timescale and to ensure the standard fully meets the needs of its users.
- The standard should initially be published as a Technical Report Type 2 to enable the developing standard to stabilize during a period of the user trials, prior to its issuing as a full International Standard.

[2]ISO/IEC JTC1 has a variety of paths to develop international standards. One of them is the path of Technical Report (TR). The TR follows a series of stages such as NP (New Proposal), WD (Working Draft), PDTR (Proposed Draft Technical Report), DTR (Draft Technical Report) and TR (Technical Report). PDTR documents are balloted at the subcommittee level. TR2 is a publication of an emerging standard in its last stage when the subject in question is still under technical development or where for any other reason there is the possibility of an agreement at some time in the future (ISO/IEC JTC1 Directives 1999). At the time of writing, ISO/IEC 15504 is currently at the stage of TR2.

Copyright © 2001 John Wiley & Sons, Ltd.

The SPICE Project Organization was subsequently established in June 1993 as an international collaborative effort to:

- Assist the standardization project in its preparatory stage to develop initial working drafts;
- Undertake user trials in order to gain early experience data that will form a basis for revision of the published technical reports prior to being reviewed as a full International Standard;
- Create market awareness and take-up of the evolving standards.

The SPICE Project adopted a fast development route for the standard and it is undertaking a set of trials to validate the standard against the goals and requirements defined at the start of the SPICE Project (ISO/IEC WG10/N017R 1993), and to verify the consistency and usability of its component products. The SPICE Project set up a trials team in order to organize the trials and to collect and analyze the data collected from the trials. The team has met regularly since 1993 in parallel with JTC1/SC7/WG10[3]. In particular, the team is responsible for

- designing data collection procedures to collect data from the trial assessments;
- maintaining the data in the trials database;
- monitoring the progress of the trials;
- distributing the data to the analysis team in a non-attributable form;
- performing analysis on the data in line with the goals of the various phases of the trials.

The original trials plan organized the SPICE Trials into three broad phases as follows:

- *Phase 1* took place in 1995 and its goals were to validate the design decisions inherent in the initial document set as well as to test the usability of the core product documents (SPICE version 1).
- *Phase 2* took place between September 1996 and June 1998 and was based on the PDTR (Proposed Draft Technical Report) version of the emerging ISO/IEC 15504 standard. In addition to evaluating the complete document set and design decisions, its objectives include providing guidance for applying the emerging standard most effectively. Phase 2 of the SPICE Trials evaluates the ISO/IEC PDTR 15504 documents (SPICE version 2).
- *Phase 3* began in July 1998 and is expected to continue until ISO/IEC 15504 becomes a full International Standard. It is based on the TR (Technical Report) version of the ISO/IEC 15504 standard. Its goal is to validate the overall goals and requirements of the standard.

2.2. ISO/IEC 15504 Compatibility and Conformance

The PDTR and TR versions of the ISO/IEC 15504 standard comprise nine documents (known as *Parts*). The details of the parts and the relations between them are described in Appendix A. ISO/IEC 15504: Part 2 (*A reference model for processes and process capability*) provides the requirements for compatibility in SPA. The requirements for compatibility enable the comparison of outputs from assessments that use different models and/or methods. There are also ISO/IEC 15504 requirements that pertain to the actual conduct (planning as well as performance) of an assessment. If an assessment is conducted to satisfy the requirements in ISO/IEC 15504: Part 3 (*Performing an assessment*), then the assessment is said to be ISO/IEC 14404 conformant. One of the requirements is that a ISO/IEC 15504 compatible assessment model is used.

Any assessment model[4] fulfilling the requirements is claimed to be ISO/IEC 15504 conformant. For example, Bootstrap 3.0 claims compliance with ISO/IEC 15504 (Kuvaja 1999) and the CMMI[SM] (Capability Maturity Model Integration) product suit is claimed to be 'consistent and compatible' with the standard (CMMI 1999, CMMI 2000). A mapping between processes defined in ISO/IEC 15504 and SW-CMM® (Paulk *et al*. 1993) is addressed by Paulk (1998). El-Emam and Garro (2000) estimated that

[3]Working Group 10 of Subcommittee 7 (Software Engineering Standardization) under a Joint Technical Committee 1 for ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission). WG10 is working for development of standards and guidelines covering methods, practices and application of process assessment in software product procurement, development, delivery, operation, maintenance and related service support.

[4]In ISO/IEC 15504, the assessment model refers to the model used as the basis for an assessment. Part 5 is an assessment model as an exemplar model. SEI calls the SW-CMM as a reference model. At the point of ISO/IEC 15504, the SW-CMM and Bootstrap are referred to as assessment models. CMM and Capability Maturity Model are registered in the U.S. Patent and Trademark Office. [SM]CMMI is a service mark of Carnegie Mellon University.

Copyright © 2001 John Wiley & Sons, Ltd.

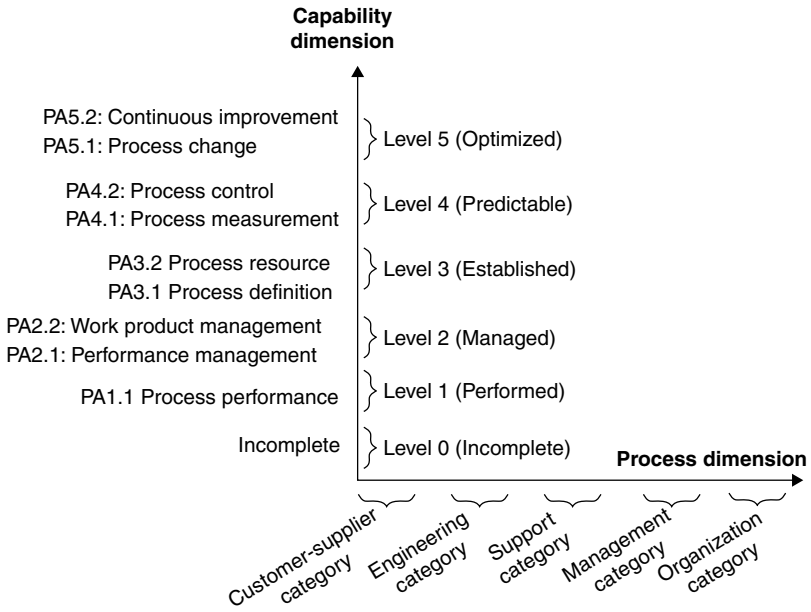

Figure 1. Two-dimensional architecture of ISO/IEC 15504, where PA denotes Process Attribute

over 1250 SPICE-based trial assessments took place during Phase 2 of SPICE Trials.

2.3. ISO/IEC 15504 Overview

2.3.1. Two-dimensional Architecture

The architecture of the published ISO/IEC 15504 standard consists of both process and capability dimensions. Figure 1 shows the structure of the process and capability dimensions. The process dimension in the PDTR version of the model consists of 29 processes classified into five process categories known as *Customer-Supplier*, *Engineering*, *Support*, *Management*, and *Organization* process categories. The capability dimension comprises six capability levels[5] ranging from 0 to 5. The greater the level, the greater the process capability achieved.

2.3.1.1. Process Dimension The process dimension is composed of five process categories as follows (see Table B1 in Appendix B for the processes in each category):

- *Customer Supplier process category (CUS)*: processes that directly impact the customer, support development and transition of the software to the customer, and provide for its correct operation and use.
- *Engineering process category (ENG)*: processes that directly specify, implement, or maintain a system and software product and its user documentation.
- *Support process category (SUP)*: processes that may be employed by any of the other processes (including other supporting processes) at various points in the software lifecycle.
- *Management process category (MAN)*: processes that contain generic practices that may be used by those who manage any type of project or process within a software lifecycle.
- *Organization process category (ORG)*: processes that establish business goals of the organization and develop processes, products and resource assets which, when used by the projects in the organization, will help the organization achieve its business goals.

2.3.1.2. Capability Dimension As shown in Figure 1, the capability dimension comprises six capability levels ranging from 0 to 5. The capability levels are determined by measuring the process attributes

[5]The model with the ordered capability is called a continuous model, while the model with the ordered processes is referred to as a stage model. Thus, ISO/IEC 15504 is a continuous model while the SW-CMM is a stage model. However, some processes in ISO/IEC 15504 have strong relationship with the capability level. Thus, this definition is not definitive.

Copyright © 2001 John Wiley & Sons, Ltd.

(PAs) (see Table B2 in Appendix B for the PAs for each capability level). The capability dimension is summarized as follows:

- *Level 0, Incomplete*. There is a general failure to attain the purpose of the process. There are little or no easily identifiable work products or outputs of the process.
- *Level 1, Performed*. The purpose of the process is generally achieved. The achievement may not be rigorously planned and tracked. There are identifiable work products for the process, and these testify to the achievement of the purpose.
- *Level 2, Managed*. The process delivers work products according to specified procedures and is planned and tracked. Work products conform to specified standards and requirements.
- *Level 3, Established*. The defined process is performed and managed based upon good software engineering principles. Individual implementations of the process use approved, tailored versions of standard, documented processes to achieve the process outcomes. The resources necessary to establish the process definition are also in place.
- *Level 4, Predictable*. The defined process is performed consistently in practice within control limits to achieve its process goals. Detailed measures of performance are collected and analyzed. This leads to a quantitative understanding of process capability and an improved ability to predict and manage performance. Performance is quantitatively managed. The quality of work products is quantitatively known.
- *Level 5, Optimizing*. Process performance is optimized to meet current and future business needs, and the process achieves repeatability in meeting its defined business goals. Performance of quantitative process effectiveness and efficiency goals (targets) for performance are established, based on the business goals of the organization. Continuous process monitoring against these goals is enabled by obtaining quantitative feedback and improvement is achieved by analysis of the results.

2.3.2. Capability Level Determination

An ISO/IEC 15504 assessment is applied to an organizational unit (OU) (El-Emam *et al*. 1998a). An OU is the whole or the part of an organization that owns and supports the software process. During an assessment, an organization can cover only the subset of processes that are relevant to its business objectives. In most cases, it is not necessary to assess all of the processes in the process dimension. In the PDTR version of 15504, the object that is rated is the process instance. A process instance is defined to be a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner (El-Emam *et al*. 1998a).

The capability level of each process instance is determined by rating process attributes. For example, to determine whether a process has achieved capability level 1 or not, it is necessary to determine the rating achieved by PA1.1 (Process performance attribute). A process that fails to achieve capability level 1 is at capability level 0. Capability levels 2–5 each have two process attributes associated with them as shown in Table B2. A more detailed description of the attributes can be found in ISO/IEC 15504: Parts 2 and 5.

As seen in Table 1, each process attribute is measured by an ordinal rating F (*Fully*), L (*Largely*), P (*Partially*), or N (*Not achieved*) that represents the extent of achievement of the attribute as defined in ISO/IEC 15504: Part 2. A process instance is defined to be at capability level k if all process attributes below level k satisfy the rating F and the level k attribute(s) are rated as F or L. This rating scheme can be depicted as in Table 2. As an example, for a

Table 1. The rating scale of the process attributes

Acronym	Achievement of the defined attribute
N (Not achieved)	0% to 15%: there is little or no evidence of achievement of the defined attribute in the assessed process.
P (Partially achieved)	16% to 50%: there is evidence of a sound systematic approach to and achievement of the defined attribute in the accessed process. Some aspects of achievement may be unpredictable.
L (Largely achieved)	51% to 85%: there is evidence of a sound systematic approach to and significant achievement of the defined attribute in the accessed process. Performance of the process may vary in some areas or work units.
F (Fully achieved)	86% to 100%: there is evidence of a complete and systematic approach to and full achievement of the defined attribute in the assessed process. No significant weaknesses exist across the defined organization unit.

Copyright © 2001 John Wiley & Sons, Ltd.

Table 2. Capability level ratings

Process attributes	Level 1 (Performed)	Level 2 (Managed)	Level 3 (Established)	Level 4 (Predictable)	Level 5 (Optimizing)
PA1.1 (Process performance)	L or F	F	F	F	F
PA2.1 (Performance management)		L or F	F	F	F
PA2.2 (Work product management)		L or F	F	F	F
PA3.1 (Process definition)			L or F	F	F
PA3.3 (Process resource)			L or F	F	F
PA4.1 (Process measurement)				L or F	F
PA4.2 (Process control)				L or F	F
PA5.1 (Process change)					L or F
PA5.2 (Continuous improvement)					L or F

process instance to be at capability level 3, it requires F ratings for PA1.1 (Process performance), PA2.1 (Performance management), PA2.2 (Work product management) and F or L rating for PA3.1 (Process definition) and PA3.2 (Process resource).

3. DATA

3.1. Data Collection

Phase 2 of the SPICE Trials used the regional structure defined for the project as a whole, which divides the world into five regional trials centers (RTCs), namely Canada (including Latin America), Europe (including South Africa), North Asia Pacific (centered on Japan and including Korea), South Asia Pacific (centered on Australia and including Singapore) and the USA. At an earlier stage of the project there were only four RTCs, North Asia Pacific and South Asia Pacific being one RTC. At the country or state level, local trials coordinators (LTCs) liaised with the assessors and OUs to ensure assessors' qualifications, to make the questionnaires available, to answer queries about the questionnaires, and to ensure the timely collection of data. There were 26 such coordinators worldwide during the second phase of the SPICE Trials.

The data set submitted to the international trials coordinator (ITC) for each trial included the ratings data from each assessment and answers to a set of questionnaires that followed each assessment. The questionnaires, which concern the assessment, the OU, the project etc., were completed by lead assessors and OUs. During the Phase 2 trials, there were 70 assessments of 44 organizations from the five regions as shown in Figure 2: Europe (24 trials); South Asia Pacific (34 trials); North Asia Pacific (10 trials); USA (1 trial), and Canada/Mexico (1 trial)

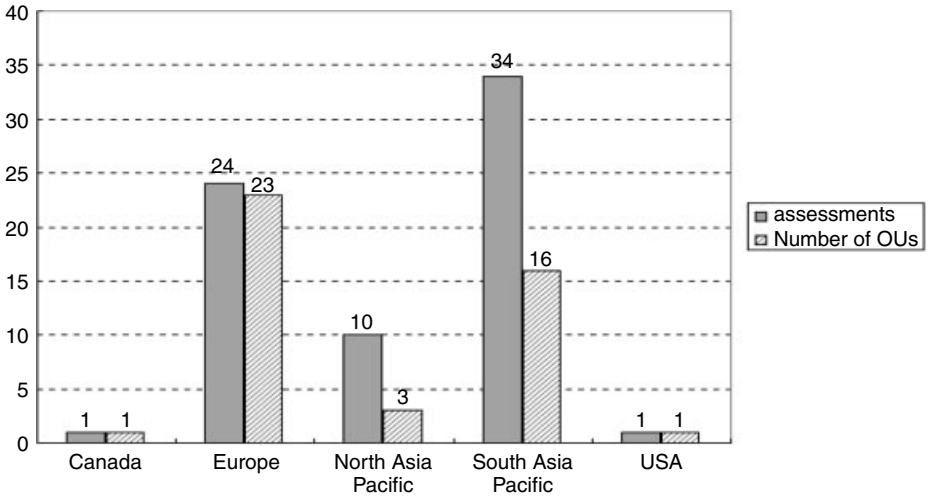

Figure 2. Assessments and OUs in region

Copyright © 2001 John Wiley & Sons, Ltd.

(Hunter 1998; El-Emam and Birk 2000a and 2000b). Since more than one assessment occurred in some OUs, the number of OUs was less than the number of assessments.

Sixteen of the 44 OUs were concerned with the production of software or other IT products or services. Using the definition of a small software organization that has also been used in the European SPIRE project (less than or equal to 50 IT staff) (Sanders 1998), we find that 52% (23/44) of the participating OUs are small. As can be seen from these data, there was adequate variation in the sizes (both small and large) of the OUs that participated in the trials.

3.2. Handling Missing Values

It is not uncommon for data sets to have missing values. This may occur because of unit non-response (i.e. the respondent did not provide any data at all), item non-response (i.e. the respondent did not answer some questions on a questionnaire), or answering questions with the 'Don't know' response category. Of course it is important to follow-up on non-respondents and attempt to get complete data sets, but more often than not this attempt still fails to bring in complete data sets. The persistent non-response problem does not only plague data collection efforts for research purposes, but also data collection within organizations, for example, software engineers failing to provide effort forms for a particular week, or failing to log some defects during unit testing.

The traditional mechanism that has been employed in many software engineering empirical studies has been to ignore missing values. The obvious disadvantage of this approach is that it leads to a loss of degrees of freedom and statistical power in subsequent analysis. Consider the situation of a multivariate analysis, then every observation with a missing value on one of the variables would have to be discarded. A more serious disadvantage is if the non-respondents differ in substantive ways from respondents, hence resulting in misleading generalizations. The problem of missing values is rarely acknowledged in software engineering empirical work.

The SPICE Trials analysis team has, for some investigations, used multiple imputation to fill in the missing values repeatedly (Rubin 1987). Multiple imputation is the preferred approach to handling missing data problems in that it provides for proper estimates of parameters and their standard errors. However, due to the need to synchronize trials' output with the progress of ISO/IEC 15504 towards an International Standard, it was not possible to perform all analyses using imputed data sets. Therefore, some of the analyses included in this paper use imputed data sets, and some follow the more common approach of ignoring missing values. A more detailed description of how multiple imputation was used can be found in El-Emam and Birk (2000a and 2000b) and El-Emam and Jung (2001). In the following sections, we will clearly identify the cases in which multiple imputation is used.

3.3. Scale Type Assumption

According to classical measurement theory concerned with 'permissible statistics' developed by Stevens (1951), variables should be measured on an interval scale if the arithmetic mean and variance are to be computed (see also Nunnally and Bernstein 1994).

To analyze data set from the trials, ratings F, L, P, and N were converted into a numerical scale by assigning the values 4, 3, 2, 1 to them, respectively, or the capability levels were coded such that 'capability level 5' was 5, down to 'capability level 0', coded 0. El-Emam and Birk (2000a and 2000b) state that the coding scheme for process capability lies between ordinal and interval level measurement. However, in the above papers they treated capability level as being on an interval scale since capability level is a single item measure that is treated as if it is interval in many instances. Furthermore, the use of non-parametric methods on non-interval scale data would exclude much useful study (Nunnally and Bernstein 1994). Many authors as well as Stevens himself noted that a useful study can be conducted even if the proscriptions are violated (Briand *et al*. 1996, Gardner 1975, Stevens 1951, Velleman and Wilkinson 1993). A detailed discussion of the scale type issue for process capability is given by El-Emam and Birk (2000a and 2000b).

3.4. Summary of the Phase 2 Trial Assessments

The Phase 2 trials collected data from 70 assessments involving 44 different organizations, 169

Copyright © 2001 John Wiley & Sons, Ltd.

projects and 691 process instances. The most common primary business sectors in which the organizations were involved were defense, IT products and services and software development. The assessments involved both large and small organizations and covered all the processes in the reference model. The median number of process instances per assessment was 6.5. The most costly activity during an assessment is the collection of evidence (47% of effort) and the least costly is the final presentation (3% of effort). The assessors involved had a broad range of experience with many having participated in Phase 1 of the trials, used the SW-CMM, and/or been involved in ISO 9001 audits (mainly in the context of the TickIT scheme). Almost all of the competent assessors (93%) received assessment training, in most cases in the context of ISO/IEC 15504.

4. DESCRIPTIVE SUMMARY OF PHASE 2 TRIALS DATA

4.1. Summary of the Projects Involved in Assessments

More than one project may have been assessed in a single assessment. We had data from the 169 projects involved in the Phase 2 SPICE Trials. The number of projects per trial is shown in Figure 3. It is evident that most assessments involved only one project. However, some covered up to 26 projects in a single assessment.

4.2. Process Coverage

In total, 691 process instances were rated in Phase 2 of the SPICE Trials. The magnitude of assessed process categories is ordered accordingly: 'Engineering' (224 process instances); 'Support' (177 process instances); 'Management' (126 process instances); 'Customer-supplier' (90 process instances) and 'Organization' (74 process instances). Presumably this was because for the organizations that participated in the trials, these ordered process categories tended to be the most important for their business. In general, ENG.2 (Develop software requirements; 56 process instances) and MAN.1 (Manage the project; 63 process instances) were assessed significantly more often than the other processes defined in the reference model. The least rated process was CUS.1 (Acquire software; 5 process instances). Its low frequency of rating is likely due to many organizations not acquiring software.

4.3. Rating and Profile Analysis

For each of the 691 individual process instances assessed, ratings were recorded for each of the attributes. The total numbers of process instances,

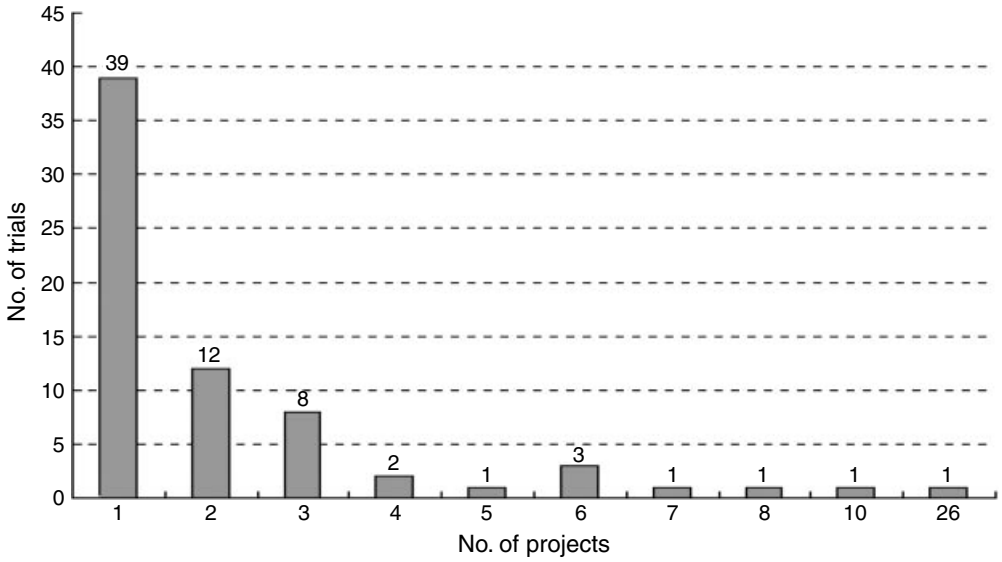

Figure 3. Number of projects per trial

Copyright © 2001 John Wiley & Sons, Ltd.

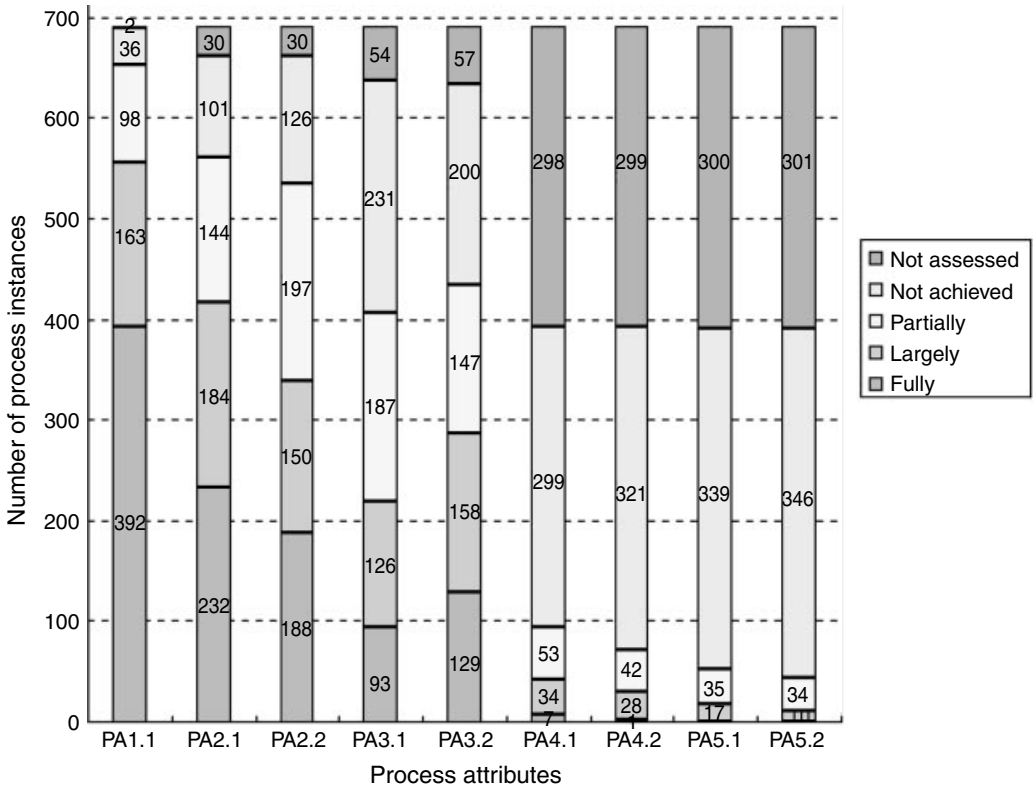

Figure 4. Frequency of process instance ratings for each process attribute

Table 3. Frequency of capability levels by process category across all process instances

Capability level	CUS	ENG	SUP	MAN	ORG	Distribution of capability level (%)
Level 0	11	23	46	35	21	19.68
Level 1	53	88	62	53	39	42.69
Level 2	20	65	32	26	9	22.00
Level 3	6	45	31	10	4	13.89
Level 4	0	3	6	2	1	1.74
Level 5	0	0	0	0	0	0

over all the trial assessments, which were rated at each capability level are shown in Figure 4. Notice that, as expected, the attributes corresponding to the higher capability levels less often receive the higher ratings than those corresponding to the lower levels. Less obvious, but worth noting, is that of the two attributes at level two, PA2.1 (Performance management) is more often highly rated than PA2.2 (Work product management) and of the two attributes at level three, PA3.2 (Process resource) is more often highly rated than PA3.1 (Process definition). At level four, it is PA4.1 (Process measurement) that seems to be more often highly rated than PA4.2 (Process control).

Table 3 shows the numbers of process instances at each capability level for each category as well as the frequency of capability during the Phase 2 SPICE Trials.

5. WHY DO ORGANIZATIONS PERFORM SPA

This section summarizes the results of a study that tried to answer the following question from

Copyright © 2001 John Wiley & Sons, Ltd.

the Phase 2 data: 'why do organizations perform SPAs?'. This study was conducted by utilizing multiple imputation. More details of the study can be found in SPICE Trials (1999) and El-Emam and Goldenson (2000).

SPA has two principal contexts for its use: capability determination and process improvement (ISO/IEC 15504: Part 1 1998). Process capability determination is concerned with analyzing the assessed capability of selected processes against a target process capability profile in order to identify the risks involved in undertaking a project using the selected processes. The proposed capability may be based on the results of relevant previous process assessments, or may be based on an assessment carried out for the purpose of establishing the proposed capability.

Within a process improvement context, SPA provides the means of characterizing the current practice within an organizational unit in terms of the capability of the assessed processes. Analysis of the results in the light of the organization's business needs identifies the strengths, weaknesses and risks inherent in the processes. This, in turn, leads to the ability to determine whether the processes are effective in achieving their goals, and to identify significant causes of poor quality, as well as overruns in time or cost. These provide the drivers for prioritizing improvements to processes. In addition, Dunaway and Masters (1996) state that SPA for SPI tends to support, enable and encourage an organization's commitment to SPI by creating a climate for change within an organization.

The Phase 2 SPICE Trials collected data[6] from assessment sponsors on the degree of importance to their organization of each of 12 reasons for performing the assessments. Table 4 gives the twelve reasons arranged in order of importance according to answers given by the sponsors.

The first nine reasons were statistically significant.[7]. The most important reason, 'Establish baseline', implies that the sponsors regarded assessment as an important measurement procedure. The next five reasons (numbered by 2 to 6) imply that the sponsors considered SPA as a basis for SPI. Reasons 7–9 denote 'buy-in' for SPI. The respondents were indifferent to the next two reasons, 10 and 11 (i.e. the respondents thought the reasons were neither important nor not important for conducting an assessment). The last reason (competitive/marketing pressure to demonstrate process capability) was not seen as a reason for performing assessments. El-Emam and Goldenson (2000) suggest this as 'a reflection of the fact that the OUs in the SPICE Trials are innovators and early adopters, and therefore their priorities may be different from the general population of organizations.' As an example of market pressure, some of the larger procurers of software systems regard the results of SPA as an important element in software supplier selection (Coallier *et al*. 1999, Saiedian and Kuzara 1995, Rugg 1993).

Table 4. The order of importance to reasons for performing SPA

	Reasons
1.	Establish baseline and/or track the organization's process improvement
2.	Improve efficiency
3.	Establish best practices to guide organizational process improvement
4.	Establish project baselines and/or track projects/process improvement
5.	Improve customer service
6.	Customer demand to improve process capability
7.	Generate management support and buy-in for software process improvement
8.	Generate technical staff support and buy-in for software process improvement
9.	Improve reliability of products
10.	Improve reliability of services in supporting products
11.	Gain market advantage
12.	Competitive/marketing pressure to demonstrate process capability

[6]The wording of the question was 'To what extent did the following represent important reasons for performing a software process assessment?'. The response categories were: 'Don't Know'(treated as missing), 'Very Important' (coded as 1), 'Important' (coded as 2), 'Somewhat Important' (coded as 3), 'Not Very Important' (coded as 4), 'Not At All Important' (coded as 5).

[7]There are two alternatives for identifying the importance for each one of the above twelve reasons. The first is to dichotomize the five-point scale into 'Important' and 'Not Important' categories. This method encountered difficulty in deciding which category the response 'Somewhat Important' should be in. Thus, the second alternative, and the one that we opted for, was to use the mean and confidence interval. If the confidence interval covers the value of three, then there is evidence, at a two-tailed α level of 0.1, that the mean response for that question is not different from 3, i.e. insignificant or indifference. Detailed statistical techniques can be found in El-Emam and Goldenson (2000) and SPICE Trials (1999).

Copyright © 2001 John Wiley & Sons, Ltd.

The above results are consistent with the two contexts for performing SPA, i.e. identifying strengths, weaknesses and risks inherent in the processes (capability determination) and providing the drivers for prioritizing improvements to processes (process improvement). There were no differences between small and large organizations in the reasons for performing SPA, where the definition of a 'small' organization is one with no more than 50 IT staff (Sanders 1998).

6. EVALUATING THE RELIABILITY OF PROCESS CAPABILITY MEASURES

There are two aspects to the reliability of SPAs: internal consistency and interrater agreement. This section covers internal consistency (reliability) studies concerning process capability as defined by ISO/IEC 15504. Interrater agreement is the subject of Section 7. Results of this section are based on the final data set of the Phase 2 SPICE Trials. In addition, a reliability study conducted by Jung (2001b) for the process capability measures during Phase 3 of the trials is described to support the results of Phase 2 analyses. There is no missing data as far as the ratings of process attributes is concerned. Initial study of internal consistency can be found in El-Emam (1998) and El-Emam and Goldenson (2000).

6.1. Reliability of Process Capability Measures

Since SPA involves a subjective measurement procedure, the reliability of this procedure is vital in order to have confidence in the assessment results. Reliability is defined as the extent to which the same measurement procedure yields the same results on repeated trials (Carmines and Zeller 1979). Lack of reliability is caused by random measurement error.

A basic concept for discussing the reliability of measurements is the notion of a *construct*. A construct is a meaningful conceptual object that is neither directly measurable nor observable. However, a set of items (variables, indicators) is posited to reflect an underlying construct. For example, process capability is a construct. The nine process attributes of ISO/IEC 15504 correspond to items that measure the process capability level. Thus, process capability level can be indirectly measured by considering the values of the nine process attributes. We can say that ISO/IEC 15504 has a nine-item (or PA) instrument to measure internal consistency of capability level.

Internal consistency is affected by ambiguities in wording and inconsistencies in the interpretation of wording by respondents. Readers can find internal reliability studies of the capability dimension for the SPICE Version 1.0 and ISO/IEC PDTR 15504 as well as the 1987 SEI maturity questionnaire (Fusaro *et al*. 1998).

6.2. Estimating Internal Consistency

6.2.1. Definition

Though there is a variety of internal reliability estimation methods such as test–retest, alternative-form method, split-halves method and internal consistency method (Carmines and Zeller 1979), the most commonly used method in software engineering and management information systems is the internal consistency method that utilizes the Cronbach alpha coefficient (Cronbach 1951). The Cronbach alpha coefficient was used for the internal consistency of the 1987 maturity questionnaire (Humphrey and Curtis 1991, Humphrey *et al*. 1991), of the organizational maturity instrument (El-Emam and Madhavji 1995) and of the ISO/IEC PDTR 15504 capability dimension (El-Emam *et al*. 1998a, Fusaro *et al*. 1998, SPICE Trials 1998).

The type of scale used in the most common assessment instruments is a summative one. This means that the individual ratings for each item are summed up to produce an overall rating (score). One property of the covariance matrix for a summative rating is that the sum of all terms in the matrix gives exactly the variance of the scale as a whole.

The variability in a set of item scores is attributed to one of the following two circumstances: (a) actual variation across the organizations in capability (i.e. true variation in the construct being measured) which can be considered as the signal component of the variance; (b) error which can be considered as the noise component of the variance. Computing the Cronbach alpha coefficient involves partitioning the total variance into signal and noise. The proportion of total variation that is signal equals the alpha coefficient.

The signal component of variance is considered to be attributable to a common source, presumably the true score of the construct underlying the items. When capability varies across the different organizations, scores on all the items will vary with

Copyright © 2001 John Wiley & Sons, Ltd.

it because it is a cause of these scores. The error terms are the source of unique variation that each item possesses. Whereas all items share variability due to capability, no two items are considered to share any variation from the same error source. The coefficient alpha is computed by

$$\alpha = \frac{N}{(N-1)}\left[1 - \sum \sigma_i^2/\sigma_y^2\right] \text{ or } \alpha = \frac{N\bar{\rho}}{1+\bar{\rho}(N-1)}$$

where N is the number of items. σ_i^2 and σ_y^2 are a unique variation of item i and total variation, respectively. $\bar{\rho}$ is equal to the mean interitem correlation.

The Cronbach alpha coefficient varies between 0 and 1. If there is no true score but only error in the items, then the variance of the sum will be the same as the sum of variances of the individual items. Therefore, coefficient alpha will be equal to zero. If all items are perfectly reliable and measure the same thing, then coefficient alpha is equal to one (i.e. the proportion of true score in the scale is 100 per cent).

What constitutes a satisfactory level of reliability depends on how a measure is being used and how the internal consistency value is determined good or bad. In the early stages of the research on assessment instruments, reliabilities of 0.7 or higher are considered sufficient. For basic research, a value of 0.8 is acceptable. However, in applied settings where important decisions are made with respect to assessment scores, a reliability of 0.9 is the minimum that would be acceptable (Nunnally and Bernstein 1994). Since the SPICE capability dimensions are both being used in making important decisions, the minimal tolerable value of internal consistency for these instruments should be set at 0.9 (El-Emam 1998a; Fusaro *et al*. 1998).

6.2.2. An Example for Calculating Internal Consistency

To illustrate the computation of the Cronbach alpha coefficient, let us use a case with 168 process instances from 15 assessments in Korea (see Section 6.4 in detail). All of the 15 assessments were conducted only up to capability level 3. The sample variances of each of process attributes, σ_i^2, are 0.35 (PA1.1), 0.98 (PA2.1), 1.13 (PA2.2), 0.90 (PA3.1), and 0.84 (PA3.2). The sum of the sample variance, $\sum_{i=1}^{5} \sigma_i^2$, is 4.20. For the five-PA instrument, the sample variance of the sum of the five process attributes is 16.43. Then, the Cronbach alpha coefficient would be 0.93, i.e. $(5/4) \times (1-4.20/16.43)$. The value of the sample variance is computed as $\sum(x_j - \bar{x})^2/(N_{\text{OBS}} - 1)$, where the x_j's are the codified actual values of ratings for each instance j $(j = 1, \ldots, N_{\text{OBS}})$, $\bar{x}$ is the mean of all process instances and N_{OBS} is the total number of process instances.

6.3. Dimensionality

The Cronbach alpha coefficient assumes that the construct being measured is unidimensional (Carmines and Zeller 1979). As the name implies, unidimensional scaling is relevant to those situations in which it is presumed that there exists a single, fundamental dimension underlying a set of data items (McIver and Carmines 1981). In contrast to unidimensional models, multidimensional scaling implies that there is more than a single dimension that underlies a set of items.

If the ISO/IEC 15504 capability scale is multidimensional, then it would be more appropriate to compute the internal consistency coefficient for each dimension separately. For this purpose, principal components analysis (Kim and Mueller 1978) was utilized to evaluate the dimensionality of the capability.

6.4. Results of Process Capability Reliability

The final data from the Phase 2 SPICE Trials includes ratings of 691 process instances. During an assessment, it is not always the case that all of the attributes up to capability level 5 are rated. In fact, 56, 299, and 301 process instances were not rated at capability levels 3, 4 and 5, respectively. To investigate the dimensionality, a principal components analysis with varimax rotation was performed with ratings of 390 process instances that were assessed up to capability level 5. For this purpose, ratings F, L, P and N were converted into a numerical scale by assigning the values 4, 3, 2, 1 to them, respectively.

The result in Table 5 shows that there is clearly a two-dimensional structure, with the attributes from levels 1 to 3 in one dimension, and the attributes from levels 4 and 5 in the other dimension. These two dimensions are named as 'Process Implementation' and 'Quantitative Process Management', respectively (El-Emam 1998).

Copyright © 2001 John Wiley & Sons, Ltd.

Table 5. Results of principal components analysis

Process attribute	SPICE Phase 2 Trials (390 process instances assessed up to level 5)		SPICE Trials in Korea (168 process instances assessed up to level 3)
	Factor 1 Process Implementation	Factor 2 Quantitative Process Management	Factor 1 Process Implementation
PA1.1	**0.80**	−0.05	**0.85**
PA2.1	**0.89**	0.07	**0.94**
PA2.2	**0.86**	0.15	**0.94**
PA3.1	**0.78**	0.30	**0.89**
PA3.2	**0.81**	0.28	**0.85**
PA4.1	0.32	**0.76**	
PA4.2	0.19	**0.85**	
PA5.1	0.05	**0.87**	
PA5.2	0.03	**0.91**	
	74% of variation explained		80% of variation explained

For a Korea data set assessed only to level 3, a principal components analysis showed unidimensional up to capability level 3 as seen in the last column of Table 5. This supports the above result. Note that though the Phases 2 and 3 trials used different versions of ISO/IEC 15504, the PDTR and TR versions, respectively, there is no difference in the definitions of process attributes and the capability dimension between the two versions.

The Cronbach alpha coefficient assumes unidimensionality. Thus, each of the two dimensions should be treated separately (Carmines and Zeller 1979). The first dimension (Process Implementation) and the second dimension (Quantitative Process Management) consist of 5 and 4 process attributes, respectively.

Table 6 shows the Cronbach alpha coefficient for each dimension. As shown in Table 6, the Cronbach alpha coefficients are close to the threshold value of 0.9. Therefore, the internal consistency of ISO/IEC 15504 was relatively high to be usable in practice. Note that 635 process instances were accessed up to level 3 (5 process attributes).

In an initial study of internal consistency (El-Emam 1998), the Cronbach alpha coefficient for each of the two dimensions is 0.89 (312 process instances) and 0.90 (232 process instances), respectively. For a Korea data set of ISO/IEC TR 15504 capability level that was assessed only to level 3, the Cronbach alpha coefficient up to level 3 has a high value of 0.93 (Jung 2001b). This is a higher value than the Phase 2 result. This result supports the result of Phase 2 of the SPICE Trials.

Table 6. Cronbach alpha coefficients for different numbers of attributes

Attributes up to capability level 3 (5 attributes; 635 process instances)	Attributes in capability levels 4 and 5 (4 attributes; 390 process instances)
0.88	0.87

Jung and Hunter (2001c) investigated the possibility of increasing the internal consistency by the change of the current four-category scale. For this purpose, they computed the Cronbach alpha coefficient for the following two cases:

- Combining the two middle categories of the achievement scale (L and P). If confusion between these two categories affects the internal consistency, then it would be expected that the Cronbach alpha would increase when these two categories are combined. This results in a three-category scale (F, [L, P], N).
- Combining the categories at the ends of the scale (F and L, and P and N). If confusion between the F and L categories and the P and N categories influences the internal consistency, then it would be expected that the Cronbach alpha would increase when these two categories are combined. This results in a two-category scale ([F, L], [P, N]).

For the two cases, there is a clear two-dimensional structure, with the attributes from levels 1 to 3 in one dimension, and the attributes from levels 4 and 5 in the other dimension (see Jung and Hunter, 2001c). As can be seen in Table 7, the current rating scheme has the highest Cronbach alpha

Copyright © 2001 John Wiley & Sons, Ltd.

Table 7. Cronbach alpha coefficients for different numbers of attributes

	Attributes up to capability level 3 (5 attributes; $n = 635$)	Attributes in capability levels 4 and 5 (4 attributes; $n = 390$)
Current rating scheme (N, P, L, F)	0.88	0.87
Rating scheme (N, [P, L], F)	0.86	0.83
Rating scheme ([N, P], [L, F])	0.82	0.75

coefficient. This implies that the four-category scale cannot be improved in terms of internal consistency by reducing it to a three- or a two-category scale.

6.5. Final Remarks

An interesting result from this study is the determination that software process capability, as measured by ISO/IEC 15504, is a two-dimensional construct: Process Implementation (levels 1 to 3) and Quantitative Process Management (levels 4 and 5). The Cronbach alpha coefficient for each dimension is under a threshold value of 0.9. However, the result from the Korean data shows that the first dimension has high enough internal consistency to be usable in practice. In addition, the four-category scale cannot be improved in terms of internal consistency by reducing it to a three- or two-category scale. Such internal consistency studies will continue to be performed as the ISO/IEC 15504 document set evolves to an International Standard.

7. EVALUATING THE RELIABILITY OF ASSESSMENTS

This section is based on a number of interrater agreement studies on the data from SPAs in the context of SPICE Trials (El-Emam *et al*. 1996a, b, 1997, El-Emam and Marshall 1998, El-Emam 1999).

7.1. Reliability of Assessments

The second type of reliability required to give confidence in the results of SPICE assessments is external reliability or interrater agreement. Interrater agreement is the degree of agreement in the ratings given by independent assessors to the same software engineering practices. If different assessors, each satisfying the competency requirements of the ISO/IEC 15504 framework, are presented with the same evidence, they will, ideally, produce exactly the same ratings. In practice, however, the subjective nature of the ratings will make it most unlikely that there will be perfect agreement in all cases. Evaluating interrater agreement is useful for ascertaining the reliability of an assessment based on ISO/IEC 15504.

For conducting interrater agreement studies, the assessment team is divided into two or more groups. Ideally all groups should be equally competent in rating the adequacy of the process attributes. In practice, assessors in each group need only meet minimal competence requirements since this is more congruent with the manner in which the 15504 documents would be applied. Each group would be provided with the same information, and then they would perform their ratings independently. Subsequent to the independent ratings, the groups would meet to reach a consensus or final assessment team rating. General guidelines for conducting interrater agreement studies are given in Simon *et al*. (1997), El-Emam (1999) and El-Emam and Goldenson (2000).

7.2. Estimating Interrater Agreement

7.2.1. Definition

To evaluate interrater agreement, the ISO/IEC 15504 achievement ratings are treated as being on a nominal scale. For nominal scales, Cohen's Kappa (1960) is the most popular index to describe the strength of agreement using a single summary index. It compares the agreement obtained with that expected if the ratings were independent. The value of Kappa is the ratio of observed excess over chance agreement to the maximum possible excess over chance agreement.

Table 8 is a two-way contingency table showing the assessment ratings of the two groups (teams). The cell counts are denoted by n_{ij}, with $n = \sum_{(i,j)} n_{ij}$ denoting the total sample size. The cell proportions and cell counts are

$$p_{ij} = \frac{n_{ij}}{n}$$

Copyright © 2001 John Wiley & Sons, Ltd.

Table 8. Interrater agreement in an assessment

		Team 2				n_{i+}
		F	L	P	N	
Team 1	F	n_{11}	n_{12}	n_{13}	n_{14}	n_{1+}
	L	n_{21}	n_{22}	n_{23}	n_{24}	n_{2+}
	P	n_{31}	n_{32}	n_{33}	n_{34}	n_{3+}
	N	n_{41}	n_{42}	n_{43}	n_{44}	n_{4+}
	n_{+j}	n_{+1}	n_{+2}	n_{+3}	n_{+4}	n

where p_{ij} is the proportion of ratings classified in cell (i,j). Let P_{i+} and P_{+j} define as follows:

$$P_{i+} = \sum_{j=1}^{4} p_{ij} = \sum_{j=1}^{4} \frac{n_{ij}}{n} \text{ and } P_{+j} = \sum_{i=1}^{4} p_{ij} = \sum_{i=1}^{4} \frac{n_{ij}}{n}$$

Then, P_{i+} and P_{+j} are the total proportion for row i and column j, respectively.

The most straightforward approach to evaluate agreement is to consider the proportion of ratings upon which the two groups agree

$$P_o = \sum_{i=1}^{4} p_{ii}$$

However, this value includes agreement that could have occurred by chance. Thus, the use of percentage or proportion agreement is not recommended as an evaluation measure. If the assessor's ratings were at random according to the marginal proportions, then the probability of chance agreement (derived from the multiplication rule of the probability assuming independence between the two groups) equals

$$P_e = \sum_{i=1}^{4} p_{i+}p_{+i}$$

Then, Cohen's Kappa is defined by

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

The numerator, $P_o - P_e$, denotes the difference between observed agreement and by chance agreement. The denominator, $1 - P_e$, is the maximum possible excess over chance agreement.

If there is complete agreement, then $\kappa = 1$. If the observed agreement is greater than the agreement occurred by chance, then $\kappa > 0$. If the observed agreement is less than would be expected by chance, then $\kappa < 0$.

The standard version of the Kappa coefficient assumes that all disagreements are equally serious. A weighted version of Kappa allows different levels of seriousness to be attached to different levels of disagreement (Cohen 1968). The weighted version of Kappa was used in studies on the reliability of SPAs (El-Emam and Goldenson 1995; El-Emam *et al*. 1996a; Fusaro *et al*. 1997). However, thus far no weighting scheme with a substantive meaning has been developed for SPAs.

7.2.2. An Example for Calculating Kappa Value

This section illustrates an example to calculate the Kappa value. The data set came from an assessment conducted in Korea during the Phase 2 SPICE Trials. Ten assessors participated in the assessment and all of them were qualified for SPICE assessments. The ten assessors were divided into two groups for the purpose of studying interrater agreement. A detailed description can be found in SPICE Trials (1999). For the frequency in Table 9

- $P_o = \sum_{i=1}^{4} p_{ii} = (10 + 5 + 1)/20 = 0.8$.
- $P_e = \sum_{i=1}^{4} p_{i+}p_{+i} == (12 \times 10 + 5 \times 9 + 3 \times 1)/20^2 = 168/400 = 0.42$.
- Hence, the Kappa value is $\kappa = (0.8 - 0.42)/(1 - 0.42) = 0.655$.

7.3. Findings from Interrater Agreement Studies

7.3.1. Kappa Benchmark in SPA

After calculating the value of Kappa, the next question is 'how do we interpret it?' The only existing benchmarks for interpreting Kappa value came from the social science and medical studies (see El-Emam (1999) for existing Kappa benchmarks). Applying those benchmarks to SPA discipline is questionable. Therefore, El-Emam (1999) presented

Table 9. Frequency of interrater agreement

		Team 2				n_{i+}
		F	L	P	N	
Team 1	F	10	2			12
	L		5			5
	P		2	1		3
	N					
	n_{+j}	10	9	1		20

Copyright © 2001 John Wiley & Sons, Ltd.

Table 10. Kappa benchmark in ISO/IEC 15504 assessments

Kappa statistic	Strength of agreement	Percentile interpretation
≤0.44	Poor	Bottom 25%
0.44–0.62	Moderate	Bottom 50%
0.62–0.78	Substantial	Top 50%
>0.78	Excellent	Top 25%

a benchmark specific to SPAs as shown in Table 10. The value in Table 10 indicates the quartile values of Kappa coefficient from actual studies. For example, 25% of assessed process instances had Kappa values below or equal to 0.44, 25% had values greater than 0.78.

The benchmark in Table 10 can be used to decide the extent to which the reliability of new assessments is good or bad compared to assessments conducted during SPICE Trials. Such an evaluation measure can be utilized to monitor the improvement of software process assessment methods.

The benchmark was derived from the results of four interrater agreement studies from the SPICE Trials. The four were El-Emam *et al*. (1996a, b, 1997) and El-Emam and Marshall (1998). Further details of the processes assessed in the four data sets can be found in El-Emam *et al*. (1998). A detailed study for deriving Table 10 can be found in El-Emam (1999).

7.3.2. Kappa Value For Process Attributes

A study for evaluating interrater agreement for the five capability attributes is shown in Table 11. As can be seen, the ratings on all five attributes show at least moderate agreement. The data set of the study was from 40 process instances of two assessments conducted in France during Phase 2 of the SPICE Trials. The same two external assessors conducted both assessments. Both assessors met the minimal guidance stipulated in the ISO/IEC 15504 documents, in terms of experience and background. For a detailed description of the work refer to Simon *et al*. (1997).

Agreement was also found to be almost always higher than 'moderate agreement'. If it is accepted that moderate agreement is a minimum for practical usage, then these results are encouraging for users of ISO/IEC 15504.

In order to investigate possible sources of disagreement on the four-point scale, the weighted Kappa coefficient was computed for the following two cases:

- Combining the two middle categories of the achievement scale (L and P). If there is confusion between these two categories, then it would be expected that agreement would increase when these two categories are combined. This results in a three-category scale (F, [L,P], N).
- Combining the categories at the ends of the scale (F and L, and P and N). If there is confusion between the F and L categories and the P and N categories, then it would be expected that agreement would increase when these categories are combined. This results in a two-category scale ([F,L], [P,N]).

The results of this analysis are shown in Table 12. It is observed that in most cases the four-point scale provides the highest Kappa values when compared with the three- or two-category scales. The zero value for attribute 2.2 is due to the data set exhibiting

Table 12. Comparing achievement scales with different number of response categories

Process attribute	Four-category	Three-category	Two-category
PA1.1	0.78	0.59	0.78
PA 2.1	0.64	0.42	0.56
PA 2.2	0.60	0.64	0
PA 3.1	0.64	0.52	0.63
PA 3.2	0.86	0.84	0.79

Table 11. Results of interrater agreement

Process attribute	Description of attribute	Weighted Kappa value	Interpretation
PA1.1	Process performance attribute	0.78	Substantial
PA2.1	Performance management attribute	0.64	Substantial
PA2.2	Work product management attribute	0.60	Moderate
PA3.1	Process definition attribute	0.64	Substantial
PA3.2	Process resource attribute	0.86	Excellent

Copyright © 2001 John Wiley & Sons, Ltd.

very little variation when reduced to a two-category scale, and this tends to attenuate the values of Kappa. The conclusion from this table is that the four-category scale cannot be improved in terms of reliability by reducing it to a three- or a two-category scale.

8. THE VALIDITY OF PROCESS CAPABILITY MEASURES

This section is based on two studies conducted by El-Emam and Birk (2000a, b) on predictive validity of process capability measures. The results are based on multiple imputation.

As noted in Section 6, internal reliability focuses on the extent to which an empirical instrument provides consistent results across repeated measurements. Reliability is a necessary condition for any measurement instrument. An instrument must also be valid as well as reliable. Any measuring device is valid if it does what it intends to do. Validity is not related to the measuring instrument itself but the measuring instrument in relation to the purpose for which it is being used. Since ISO/IEC 15504 defines a scheme for measuring the capability of software processes, its validity must be demonstrated before one will have confidence in its use.

8.1. Predictive Validity

A basic premise of 15504 is that the quantitative score from the assessment is related to the performance[8] of an organization or project. This premise consists of two parts:

- that the practices defined in the assessment model are indeed good practices and their implementation will therefore result in good performance;
- that the quantitative assessment score is a true reflection of the extent to which these practices are implemented in the organization or project, and therefore projects or organizations with higher assessment scores are likely to perform better.

Furthermore, improving the software engineering practices according to the assessment model is expected to subsequently improve the performance. This is called the predictive validity of the process capability score. Testing this premise can be considered an evaluation of the predictive validity of the assessment measurement procedure (El-Emam and Goldenson 1995).

8.2. Evaluating Predictive Validity

8.2.1. Measurement of Performance

A previous study in Section 6 had identified the capability scale of ISO/IEC 15504 as two-dimensional: 'Process Implementation' and 'Quantitative Process Management'. The predictive validity studies are limited to the first dimension 'Process Implementation' (up to level 3) because of the dearth of observations in levels 4 and 5.

To construct a single measure of 'Process Implementation' El-Emam and Birk (2000a, b) code an F rating as 4, down to a 1 for an N rating. Subsequently, they construct an unweighted sum of the attributes at the first three levels of the capability scale. This is a common approach for the construction of summated rating scales (McIver and Carmines 1981).

The performance measures were collected through a questionnaire. The respondent to the questionnaire was the sponsor of the assessment, who should be knowledgeable about the projects that were assessed. To maintain comparability with previous studies, project performance of this study followed the definition by Goldenson and Herbsleb (1995). Their performance is 'customer satisfaction', 'ability to meet budget commitments', 'ability to meet schedule commitments', 'product quality', 'staff productivity', and 'staff morale/job satisfaction', where product quality is changed to 'ability to satisfy specified requirements' in this study.

8.2.2. Evaluation Methods

A common measure of predictive validity in general is the correlation coefficient (Nunnally and Bernstein 1994). It has also been used in the context of evaluating the predictive validity of project and organizational process capability measures (El-Emam and Madhavji 1995, McGarry *et al*. 1998).

The quantitative assessment results of ISO/IEC 15504 can be used for either supplier selection or process improvement. At the point of supplier selection, predictive validity studies use a composite measure of process capability. This means that the

[8]Carmines and Zeller (1979) use the term 'external attribute' instead of performance.

Copyright © 2001 John Wiley & Sons, Ltd.

capability of individual processes is measured, and then these individual measures are aggregated to produce an overall project or organizational measure. However, evaluation of predictive validity of individual processes is more informative than the evaluation of aggregate measures of capability. El-Emam *et al*. (1996c) addressed this point with the argument being that the capability of an individual process is unlikely to be related with all project performance measures. This is certainly true in process improvement.

El-Emam and Birk (2000a, b) evaluated the predictive validity of four processes in the Engineering category as follows:

- Develop software requirements (ENG.2);
- Develop software design (ENG.3);
- Implement software design (ENG.4);
- Integrate and test software (ENG.5).

They made use of a two-stage analysis procedure to evaluate the predictive validity. The first stage determines whether the association between 'Process Implementation' (up to level 3) of the development processes and each of the performance measures is 'clinically significant' using the Pearson correlation coefficient. This procedure indicates a magnitude that is sufficiently large. If it does, then the statistical significance of the association was exploited by utilizing an OLS (ordinary least squares) regression as follows:

$$Performance = b_0 + b_1 \times (capability\ value)$$

where the b_1 is tested whether it is different from zero. If the regression coefficient b_1 is statistically non-zero, it implies an association between performance and capability level. The model is separately applied to each of the performance measures.

In evaluating the predictive reliability, the size of the organization is considered as a context. In investigating organization size as a context factor, the process instances were divided into two groups based on whether the organization had a large or a small IT staff, where small is less than or equal to 50 IT staff. The same definition of 'small' organizations was used in a European project that provides process improvement guidance for small organizations (Sanders 1998).

8.3. Findings From Predictive Validity Study

Table 13 shows the number of OUs that assessed each of the development processes, the number of projects that were actually assessed, and the number of projects in small versus large OUs. Table 14 shows the findings from the predictive validity evaluations. In addition, this table can be used to scope an assessment according to the business objectives of an organization. For example, let us say an OU identifies budget as an important business objective, and that this OU has 100 IT staff. Then, according to Table 14, ENG.3 (Develop software design) and ENG.4 (Implement software design) should be considered for inclusion within the scope of the assessment. The rationale is that we have evidence that higher capability on these two processes is associated with ability to meet budget commitments. Therefore, if budget is important, these two processes are certainly worthy of consideration for inclusion in an assessment.

8.4. Final Remarks

The predictive validity of four processes was evaluated: ENG.2; ENG.3; ENG.4, and ENG.5. Predictive validity is the relationship between process capability and project performance, which is an underlying premise of ISO/IEC 15504. Table 14 gives the following conclusions:

- This table can also be used to link the assessment scope with business objectives.

Table 13. Number of OUs and projects that assessed each of the four software development processes

	Number of OUs	Number of projects	Number of projects in small OUs	Number of projects in large OUs
ENG.2: Develop software requirements	29	56	22	34
ENG.3: Develop software design	25	45	18	27
ENG.4: Implement software design	18	32	18	14
ENG.5: Integrate and test software	25	36	18	18

Copyright © 2001 John Wiley & Sons, Ltd.

Table 14. Summary of the predictive validity study

OU size	Performance measure	Process(es)
Small organizations	Ability to meet budget commitments	
	Ability to meet schedule commitments	Develop software design (ENG.3)
	Ability to achieve customer satisfaction	
	Ability to satisfy specified requirements	
	Staff productivity	
	Staff morale / job satisfaction	
Large organizations	Ability to meet budget commitments	Develop software design (ENG.3)
		Implement software design (ENG.4)
	Ability to meet schedule commitments	Develop software design (ENG.3)
	Ability to achieve customer satisfaction	Develop software design (ENG.3)
	Ability to satisfy specified requirements	Develop software design (ENG.3)
	Staff productivity	Develop software requirements (ENG.2)
		Integrate and test software (ENG.5)
	Staff morale/job satisfaction	Develop software design (ENG.3)

- The verisimilitude of the predictive validity premise for small organizations was supported with weak evidence. This may be an indicant that the process capability measure is not appropriate for small organizations, or that the capabilities stipulated in ISO/IEC 15504 do not necessarily improve project performance in small organizations.
- The productivity of projects in large organizations is associated with the capability of ENG.5 (the Integration and testing process). Such a relationship makes intuitive sense since this process commonly consumes large proportions of project effort.
- The association of ENG.3 and ENG.4 (the Develop and implement software design processes) and the remaining performance measures in large organizations have relatively large magnitudes, although statistical significance is only attained for ENG.2 (the Develop software design process). For ENG.4 (the Implement software design process), the sample size within that subset may have been too small.
- ENG.3 (the Develop software design process) is critical for large organizations, and its assessment and improvement can provide substantial payoff.
- Further predictive validity study is required for the remaining processes.

9. EVALUATING THE EXEMPLAR MODEL (PART 5)

This section presents a brief summary of the results of evaluating the exemplar assessment model (ISO/IEC 15504: Part 5) based on answers to questions put to assessors during Phase 2 of the SPICE Trials. A detailed study, which evaluates the percentage of supportive and critical responses, as well as the confidence interval for each of the questions asked, is described in El-Emam and Jung (2001). The results of this section are based on multiple imputation.

El-Emam and Jung (2001) show that approximately 82% of the users of ISO/IEC 15504 used the exemplar model as the basis for their assessments. This makes it important to perform systematic empirical evaluations of this model. Such evaluations should provide a substantiated basis for using the model, as well as giving the developers of ISO/IEC 15504 information as to the necessary improvements to make. In fact, one of the recurring questions during the development of ISO/IEC 15504 was 'how good is the exemplar model?'. The purpose of this section is to provide some answers to this question.

9.1. What is the Exemplar Model?

The ISO/IEC 15504 document set contains an exemplar assessment model (Part 5). One motivation for developing this model was to make it easier for organizations to use the standard immediately.

Copyright © 2001 John Wiley & Sons, Ltd.

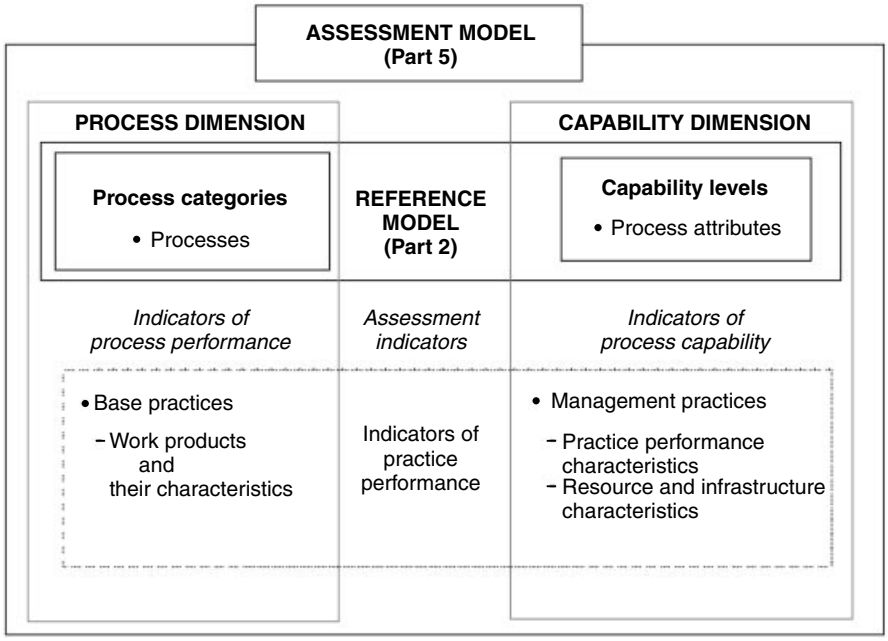

Figure 5. Relationship between the reference model and the assessment model

The basic structure of the exemplar model is identical to that of the reference model defined in Part 2. There is a one to one correspondence between the process categories, processes, process purposes, process capability levels and process attributes of the reference model and those of the exemplar model.

The exemplar assessment model expands the reference model by providing assessment indicators as shown in Figure 5. Assessment indicators are objective attributes or characteristics of a practice or work product that support an assessor's judgment of the performance or capability of an implemented process. Two different classes of indicators can be identified: indicators of process performance, and indicators of process capability. These indicator types relate, respectively, to the base practices defined for the process dimension and the management practices for the capability dimension.

9.2. Evaluating the Exemplar Model

The data that were used in this study were provided by lead assessors. They filled up a questionnaire evaluating the exemplar model after each assessment. These data were collected from the lead assessors of 57 assessments that used the model. The questionnaire was divided into the following sections:

- Use of the exemplar model.
- Usefulness and ease of use of the exemplar model.
- Meaningfulness of the rating aggregation scheme.
- Usability of the rating scale.
- Usefulness of indicators.
- Understanding of the process and capability dimensions.

The unit of analysis in our study is the assessment. Even though an individual assessor may perform more than one assessment, they were requested to answer the questionnaire once for each assessment to reflect his/her experiences during that assessment.

The objective of the analysis of the questionnaire responses was to identify the proportions of respondents who are supportive (as opposed to critical) of either the assessment model design decisions or the claim that it is usable. A supportive response is one:

- that says something positive about the assessment model, and/or
- that will *not* require any changes to the assessment model.

Copyright © 2001 John Wiley & Sons, Ltd.

9.2.1. Aggregation of Responses

We present our results in terms of the proportion of respondents who gave supportive answers. For each question, we identify the response categories that are supportive, and those that are critical. A proportion of supportive responses is then calculated. For example, assume that a question asked the respondents to express their extent of agreement to the statement 'There is sufficient detail in the assessment model to guide process improvement', and that it had the following four response categories: 'Strongly Agree'; 'Agree'; 'Disagree', and 'Strongly Disagree'. For the above question, the 'Strongly Agree' and 'Agree' responses would be considered supportive of the model, and the 'Disagree' and 'Strongly Disagree' responses would be considered to be critical of the assessment model, as shown in Table 15.

9.2.2. Interpretation of Proportions

In SPICE-related studies analyzing responses, Maclennan and Ostrolenk (1995) and Jung (2001a) used 80% of the supportive respondents as a threshold for taking action. For consistency with previous studies, 80% supportive responses are used as a boundary for interpreting results. If the estimated proportion is less than 80% then the particular issue is considered to have only moderate support.

In some cases, the percentage of supportive responses is different from 80% as a consequence of sampling variability. Therefore the confidence interval can be employed to test the hypothesis that there were less than 80% supportive respondents. The confidence interval of each question is presented in El-Emam and Jung (2001).

9.3. Analysis of Results

9.3.1. Use of the Assessment Model

Nearly all of the respondents used the exemplar assessment model as a source of indicators (95.5%). Approximately 82% of the respondents have used Part 5 intensively. Nearly half of them (45.61%) used the model to define additional indicators.

Table 15. Types and examples of response categories

Supportive responses	Critical responses
Strongly Agree	*Disagree*
Agree	*Strongly Disagree*

9.3.2. Usefulness and Ease of Use

In general, lead assessors found Part 5 both useful and easy to use. Furthermore, they were satisfied with the level of detail of the exemplar model (87.2%). However, a minority expressed some concern that they could have produced accurate judgments with less detailed evidence, suggesting that the effort on collecting evidence as stipulated by the indicators may be reduced (71.93%, but not statistically different from 80%).

This pattern of results implies that the detailed evidence collected using the exemplar model was not too much to handle, but could probably be reduced without affecting the accuracy of the ratings. This issue is particularly relevant as the size of the ISO/IEC 15504 document set has been a concern in the past (Kitson 1997) and the exemplar assessment model is the largest document of the set by far (having 38% of the total number of pages). In fact, the large size of the document set, and the exemplar model in particular, has been a recurring theme during the development of ISO/IEC 15504.

9.3.3. Meaningfulness of Rating Aggregation Scheme

Nearly all assessors (98.25%) found that the rating at the process instance level was meaningful. Approximately 88% thought that aggregation of attribute ratings across process instances was meaningful. The scheme for calculating the capability level was found to be meaningful by a substantial majority, 82.46% of respondents.

The aggregation of capability levels into a profile was meaningful to 85.97% of the assessors. Finally, a smaller but still substantial 73.68% felt that the grouping of process categories was meaningful. However, this value is not statistically different from the 80% threshold.

9.3.4. Usability of the Rating Scale

Nearly all assessors (98.25%) found that they and their assessment teams could understand the distinctions among the categories of the four point achievement scales. However, when asked a more specific question, some weaknesses in the scale appeared.

The biggest difficulty seems to be making the distinction between the L and the P response categories

Copyright © 2001 John Wiley & Sons, Ltd.

(by 43.68%), followed by distinctions between the F and L response categories (by 24.56%), and lastly the P and N categories (by 21%), where the least difficulty was encountered. Only the L and P distinction was statistically different from the 80% threshold, suggesting that action should be taken to address this confusion on the scale points.

The results suggest that ratings at the extremes of the scale are easier to make. It would be informative in future studies to determine the impact of confusion on the middle response categories to the final process capability levels assigned to process instances.

9.3.5. Usefulness of Indicators

A sizeable majority did not have difficulty relating the base (96.43%) and management (87.27%) practices to the practices within the OU. Also, a large majority (86.28%) found the process capability indicators in general to be supportive of their rating judgments.

9.3.6. Understanding of the Process and Capability Dimensions

Assessors felt confident enough in their understanding of the process categories to make consistent and repeatable judgments about the practices followed in the OU (100% supportive for CUS and ENG, and 96.77% for MAN). The only exception was the ORG process category, where 8% felt that this was not the case.

For the capability dimension almost all of the assessors were confident about their understanding up until level 3 attributes (supportive proportion: 100% for PA1.1, PA2.1, PA2.2, PA3.1 and 93.24% for PA3.2). However, the confidence level dropped for levels 4 and 5 in the perceived consistency and repeatability of their judgments (supportive proportion: 73.08% for PA4.1, 74.32% for PA4.2, 69.57% for PA5.1, 72.77% of PA5.2).

9.4. Final Remarks

The current evaluation identified a number of issues that will be subsequently investigated using more focused studies:

- Evaluate the impact of confusion between L and P ratings on the capability level since our study found that assessors had the greatest difficulty in making the distinction between these two categories.
- Evaluate the reliability of ratings at higher levels of capability, since our results showed that assessors tend to have more difficulty making ratings at levels 4 and 5. A partial solution can be found in Section 6.

10. FACTORS INFLUENCING ASSESSOR EFFORT

This section describes two studies concerned with assessor effort. The purpose of the two studies is to evaluate the factors that influence assessor effort in ISO/IEC 15504-based process assessments. Results of this section are based on studies performed by Jung and Hunter (2001b) and El-Emam *et al*. (1998b).

10.1. Assessor Effort

A major component contributing to the cost of SPA is assessor effort. Assessor effort refers to the time required by the assessors to perform SPA. Assessor effort consists of assessment input preparation, briefing the organizational unit staff, collection of evidence (e.g. reviewing documentation, interviewing assessees), production and verification of ratings, preparation of assessment results, presentation of results to management, etc. (SPICE Trials 1999). An assessment plan created by the lead assessor usually specifies the number of assessors and the estimated assessment time (in person-days or person-hours) required according to the number of processes assessed, the capability levels assessed, the project characteristics, and so on (ISO/IEC 15504: Part 3).

Assessments generally involve much effort in terms of time and cost. For process assessments using SW-CMM, Fayad and Laitinen (1997) complained of the high costs of process assessments. In addition, a survey of Dunaway *et al*. (1998) reported that more than a third of assessment team leaders expressed concern at the assessment time, as did 38% of the team members. Therefore, it is extremely important to evaluate which variables are linked to assessment effort and to identify various methods to reduce effort.

10.2. Interaction Effect

This section provides two OLS models with two explanatory variables and an interaction term. The

Copyright © 2001 John Wiley & Sons, Ltd.

two models employed different variables and measurement units. Therefore, both of the interaction effect models cannot be directly compared. However, results of both models indicate that consensus among assessors is one of the most influential factors affecting assessor effort.

10.2.1. Interaction Effect Model 1

This section provides results from Jung and Hunter (2001b). Their studies did not employ multiple imputation. Their OLS model only includes *Number of assessment* and *Consensus difficulty* as well as the nature of the effect of their interaction. Dependent variable *Assessor effort* (man-hours) for the first model was measured for each assessment during the Phase 2 SPICE Trials. The two explanatory variables are defined as follows:

- *Number of assessments* – the number of assessments by the lead assessor during the previous three years. The assessments include SPICE Phase 1 or 2, ISO 9001 (TickIT), Trillium (Coallier 1995), CMM, Bootstrap, or other type of assessments. This is a continuous variable.
- *Consensus difficulty* – the difficulty experienced by assessment teams in achieving consensus. This is a binary variable that has a value of 'High' or 'Low'.

The dependent variable, *Assessor effort*, is transformed using the natural logarithmic transformation due to non-normality, i.e. ln(*Assessor effort*). Then, an OLS regression is as follows:

$$\ln(\textit{Assessor effort}) = a_0 + a_1 \times (\textit{Consensus difficulty}) + a_2 \times (\textit{Number of Assessments}) + a_3 \times (\textit{Consensus difficulty} \times \textit{Number of Assessments})$$

The continuous variable, *Number of Assessments*, may have a strong relationship with the interaction term. This potentially introduces multicollinearity in the regression model. Therefore, the variable is centered by subtracting the mean from each raw value (Aiken and West 1991). However, the dummy variable is not centered (Jaccard *et al*. 1990). The interaction term is built using the centered variables.

For the 'Low' and 'High' values of the dummy variable in the above equation, the two equations become as follows:

$$\ln(\textit{Assessor effort})_{\textit{Consensus difficulty='Low'}} = a_0 + a_2 \times (\textit{Number of Assessments})$$

$$\ln(\textit{Assessor effort})_{\textit{Consensus difficulty='High'}} = (a_0 + a_1) + (a_2 + a_3) \times (\textit{Number of Assessments})$$

Estimated coefficients from 46 assessments are $a_0 = 3.152$, $a_1 = 1.055$, $a_2 = -0.053$ and $a_3 = 0.052$. Adj R^2 as a measure of the goodness of fit has a value of 0.601. For multicollinearity checking, condition number of 2.018 is less than a recommended upper limit of 30 (Belsley *et al*. 1980). The two equations with a moderating variable *Consensus difficulty* are generated as follows:

$$\ln(\textit{Assessor effort})_{\textit{Consensus difficulty='Low'}} = 3.152 - 0.053 \times (\textit{Number of Assessments})$$

$$\ln(\textit{Assessor effort})_{\textit{Consensus difficulty='High'}} = 4.207 - 0.001 \times (\textit{Number of Assessments})$$

The *t*-test for the slope of line *Consensus difficulty* 'Low' rejects the hypothesis of a zero slope, whereas the slope of line *Consensus difficulty* 'High' accepts the hypothesis of a zero slope. These results imply that the number of assessments by the lead assessor is a variable which serves to reduce assessor effort in achieving consensus in the case of *difficulty* 'Low', but the number of assessments is no longer a variable to reduce assessor effort in the case of *difficulty* 'High'. These conclusions can be clarified, within the range of the current data set, in Figure 6.

10.2.2. Interaction Effect Model 2

The second interaction model came from El-Emam *et al*. (1998b), where effort data (man-minutes) were gathered for each process instance. The model is to evaluate the relationship between interrater agreement (*Agree*) and consolidation[9] effort (*Effort*), considering process types such as 'Organizational' process and 'Project' process. 'Organizational' type processes are activities that would span multiple projects while 'Project' type processes could be instances of process for each project.

[9] A specific meeting is dedicated to consolidate the assessment record and to establish a consensus between the two assessors when some divergence arises for one or several attribute ratings.

Copyright © 2001 John Wiley & Sons, Ltd.

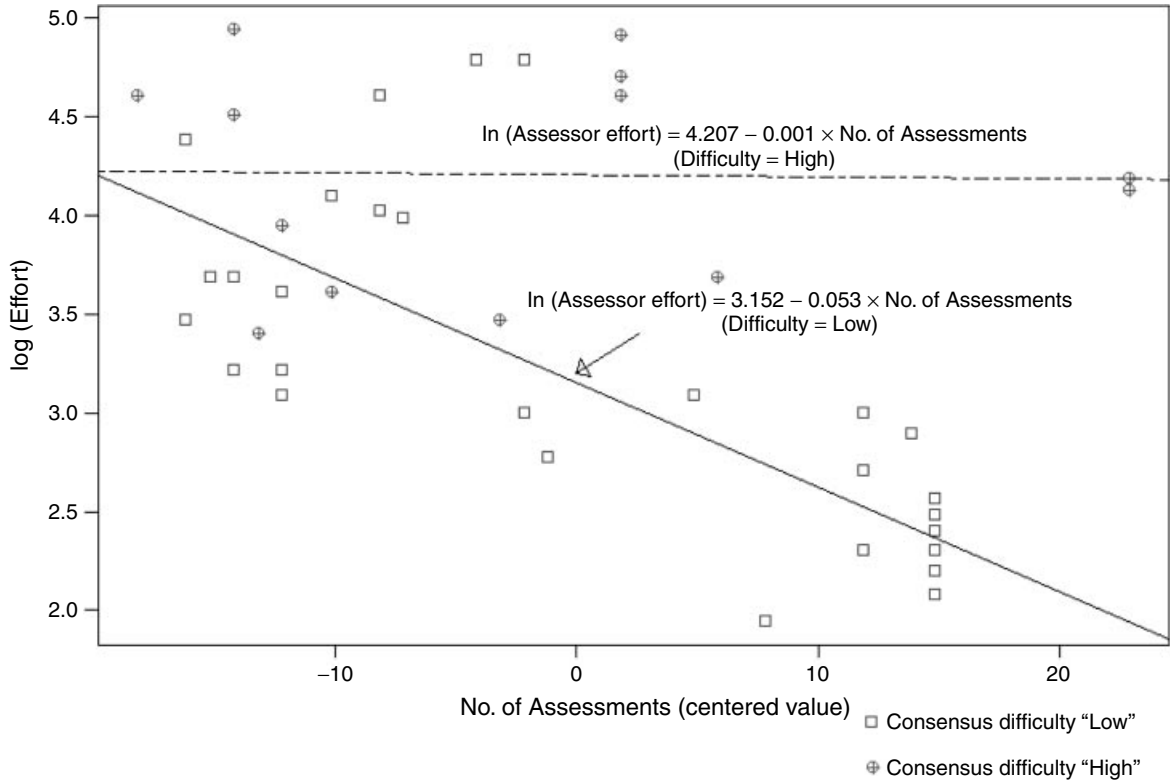

Figure 6. Interaction effect of assessor effort model

The method used for modeling the relationship was multiple ordinary least squares regression with an interaction term as follows:

$$Effort = a_0 + a_1 \times Type + a_2 \times Agree + a_3 \times (Type \times Agree)$$

Since *Type* can take only two values (organization or project), it was treated as a dummy variable in the regression model and coded 0 for 'Organization' and 1 for 'Project'. The continuous variable *Agree* is centered.

The final model has the following form with a quite good Adj R^2 value of 0.63, where all parameter are significant:

$$Effort_{TYPE=Organization} = 70 - 32 \times Agree$$
$$Effort_{TYPE=Project} = 38 + 11 \times Agree$$

The model parameters when there is centering of the continuous variable can be interpreted as follows. The parameter, a_0, is the consolidation cost for 'Organizational' processes at the mean *Agree* value for the whole sample (which is 0.713). The $(a_0 + a_1)$ value is the consolidation cost for 'Project' processes at the mean *Agree* value. In this case, the a_1 parameter represents the distance in consolidation cost between the two types of processes at the mean *Agree* value. The a_2 parameter is the slope of the line for 'Organization' type processes, and the a_3 parameter is the difference in slope between the two types of processes.

For 'Organizational' type processes higher agreement is related to a reduction in cost, while for 'Project' type processes there is no relationship. This indicates a reduction in overall costs for process assessments that include 'Organizational' type processes in their scope ensuring high agreement.

10.3. An Extended Model for Evaluating Influential Factors

10.3.1. An Extended Model

This section provides an extended model in Section 10.2.1 as follows:

Copyright © 2001 John Wiley & Sons, Ltd.

$$\ln(\textit{Assessor effort}) = b_0 + b_1(\textit{Number of PAs rated}) + b_2(\textit{Assessor experience}) + b_3(\textit{Number of Assessments}) + b_4(\textit{OU support level}) + b_5(\textit{Consensus difficulty}) + b_6(\textit{PA difficulty}) + b_7(\textit{Assessment team}) + b_8(\textit{IT size}) + b_9(\textit{ISO 9001 certification}) + b_{10}(\textit{Number of Assessments} \times \textit{Consensus difficulty}) + e$$

where the first three variables are continuous and defined as follows:

- *Number of PAs rated*: the number of process attributes rated.
- *Assessor experience*: the number of years that the lead assessor has spent in SPA.
- *Number of assessments*: defined in Section 10.2.1.

The remaining variables are binary variables. They are *OU support level, Consensus difficulty, PA difficulty, Assessment team, IT size*, and *ISO 9001 certification* and each of them has a value of the definition as follows:

$$\text{discrete variable} = \begin{cases} 1 & \text{if discrete variable has a value of 'High'} \\ 0 & \text{otherwise} \end{cases}$$

In this model, the continuous variable *Number of Assessments* is centered by subtracting the mean from each raw value to reduce potentiality of collinearity between it and the interaction term.

10.3.2. Evaluation Results and Discussion

Since the purpose of this study is to investigate factors (variables) that influence assessor effort, we are interested in positive or negative associations between assessor effort and the variables in the regression model. Thus, this is a one-tailed hypothesis.

The extended regression model meets the normality assumption test (Shapiro and Wilk 1965), the homoscedasticity (consistency of error variance) assumption (White 1980), and the condition number for checking collinearity. The Adj R^2 is high enough to derive a conclusion. As seen in Table 16, all variables except two variables (*Assessor experience, OU support level*) influence assessor effort at $\alpha = 5\%$. Our findings in the study provide several insights for assessors and OUs for SPA.

As we expected, assessor effort is positively related to the number of process attributes rated and it is negatively related to the number of assessments conducted by lead assessor. The latter implies that increased assessment experience of lead assessor leads to enhancement of assessment skills such as assessment progress control, interviewing assessees, documentation preparation etc. Rout *et al*. (2000) emphasized the number of assessments as a method to reduce assessment effort.

Assessor effort is positively related to the difficulty experienced by assessment teams in achieving consensus. Since process assessment is the collaborative work of a team of assessors, teams that experienced difficulty in achieving consensus would also have experience in increased assessment time.

Assessor effort is positively related to the difficulty that assessment teams have in determining PA ratings from the raw data concerned with base practices, management practices and work products. This problem was addressed in El-Emam and Jung (2001) in detail.

Some of the Phase 2 trials had more than one assessment team for the same processes in order to investigate issues such as interrater agreement (El-Emam 1999). An assessment team is therefore either single or multiple. A multiple assessment team took more assessment effort than a single team.

Our analysis revealed that an assessment for OUs of large IT staff size required less assessor effort than for OUs of small IT staff size. In the Phase 2 SPICE Trials, the number of PAs rated in OUs with large IT staff is smaller than that rated in OUs with small IT staff. In addition, assessor effort for a single process attribute in OUs with large IT staff is smaller than that in OUs with small IT staff. Those effects are revealed in the negative regression coefficient. Assessment for ISO 9001 certified OU took less assessor effort than that for non-ISO 9001 certified OU.

10.4. Final Remarks

Some findings in this section allow for intuitive reasoning. The more the number of process attributes

Copyright © 2001 John Wiley & Sons, Ltd.

Table 16. Regression results for assessor effort

Variables	Parameter	Coefficient	Standard error	One-sided p-value
Intercept	b_0	3.113	0.242	0.000
Number of PAs rated	b_1	0.005	0.002	0.004
Assessor experience	b_2	−0.001	0.028	0.482
Number of assessments	b_3	−0.057	0.009	0.000
OU support level (high = 1, low = 0)	b_4	0.284	0.183	0.065
Consensus difficulty (high = 1, low = 0)	b_5	0.848	0.165	0.000
PA difficulty (high = 1, low = 0)	b_6	0.320	0.159	0.026
Assessment team (multiple = 1, single = 0)	b_7	0.680	0.242	0.004
IT size (large = 1, small = 0)	b_8	−0.523	0.170	0.002
ISO 9001 certification (certified = 1, non-certified = 0)	b_9	−0.385	0.175	0.017
Number of assessments × Consensus difficulty	b_{10}	0.071	0.014	0.000
Adj R^2		0.77		
F-test (model)		15.79 ($p < 0.001$)		
Condition number		9.60		

rated, the more the assessor effort. The variable denoting difficulty to determine PA ratings, *PA difficulty*, is compatible with a previous study. It is reasonable that a multiple assessment team took more effort than a single assessment team. In ISO 9001 certified OUs, experience of assessments reduced the assessor effort.

The nine-variable OLS model has an Adj R^2 value of 0.77, whereas the interaction model with *Consensus difficulty* and *Number of Assessments* gives an Adj R^2 value of 0.601. This is a quite a good fit with two variables. In the interaction model, lead assessor experience in assessment was revealed to reduce assessor effort in the case of no consensus problems, but is ineffective in reducing assessor effort in consensus difficulty.

Consensus difficulty among assessors is related to a variety of aspects. El-Emam *et al*. (1998b) states that the consensus problem may be due to one or a combination of reasons: assessors' lack of experience, lack of knowledge of the assessment model; lack of knowledge of the capability scale, or lack of knowledge of the OU and its business. In addition, the problem may also be caused by ambiguities in the assessment model and the capability measurement scale, and weakness in the assessment model itself.

Based on analysis of the Phase 2 SPICE Trials data, El-Emam and Jung (2001) reported that the difficulties expressed by competent assessors in understanding the boundary between F and L, between L and P, and between P and N were 24.56%, 43.96% and 21.05%, respectively. Such difficulties may indicate one reason in the consensus problem that can be partially resolved by re-investigating the evidence and collecting further information. However, for situations in which this does not bring about a consensus among assessors, Jung (2001a) proposed a systemic approach based on AHP (analytic hierarchy process) to solve boundary problems.

11. EMPIRICAL COMPARISON OF ISO/IEC 15504 AND ISO 9001

This section describes an empirically based comparative study between ISO/IEC 15504 and ISO 9001. A more detailed description is provided in Jung and Hunter (2001a) and SPICE Trials (1999).

11.1. ISO 9001 and ISO/IEC 15504

ISO 9001 (1997) contains 20 clauses that collectively provide the minimum requirements for a quality management system for use in software development and maintenance, as well as in other industries. Satisfaction of all the requirements can lead to ISO 9001 certification. ISO 9000-3 (1997) contains software specific guidelines for the use of ISO 9001.

Although ISO 9001 and ISO/IEC 15504 have different origins, i.e. ISO 9001 is a generic standard for quality management and assurance while 15504 was created solely for SPA, capability determination and process improvement, the two standards are intuitively similar, as has been shown in a brief

Copyright © 2001 John Wiley & Sons, Ltd.

comparative study of the two standards (Hailey 1998). To the best of our knowledge, there are no detailed studies which assess the degree of similarity between the two standards as is witnessed in a comparative study of ISO 9001 and SW-CMM (Paulk *et al*. 1993) conducted by Paulk (1995). In the study, Paulk attempted to answer questions such as 'At what level in the CMM would an ISO 9001 compliant organization be?' and 'Can a CMM level 2 (or 3) organization be considered compliant with ISO 9001?'.

This section provides empirical answers to the following questions relating to ISO 9001 and ISO/IEC 15504:

- At what ISO/IEC 15504 capability level would one expect an ISO 9001 certified organization's processes to be?
- Is there any significant difference in the SPICE capability levels achieved by the processes of ISO 9001 certified organizations and those of non-ISO 9001 certified organizations?

11.2. Capability Levels of SPICE Processes in ISO 9001 Certified Organizations

In this section we try to answer the question 'At what SPICE capability level would one expect an ISO 9001 certified organization's processes to be?'

To analyze our data set, the capability levels were coded such that 'capability level 5' was 5, down to 'capability level 0', coded 0. The third column in Table 17 shows the average capability level of each of the 29 ISO/IEC 15504 processes for the ISO 9001 certified organizations. The average capability level for each SPICE process except CUS.1 (for which there is very little data) lies between 1 and 2.3. Since different samples will produce different values for the average capability level of a process, this study provides a confidence interval of the true value of the average capability level of each process. Figure 7 shows the 95% bootstrap confidence interval[10] of the true mean capability level for each of 15 ISO/IEC 15504 processes of the ISO 9001 certified organizations (the 14 processes with small sample size, less than nine, are not displayed). As an example, in the case of the confidence interval of CUS.2, we can say, with a confidence of 95%, that the mean capability level is in [1.533, 2.267]. Note that the average capability values should be considered conservative because some assessments did not perform assessments beyond capability level 3.

From Table 17, the processes that have average capability level greater than or equal to 2 are ENG.1 (Develop system requirements and design), ENG.6 (Integrate and test system), ENG.7 (Maintain system and software), SUP.2 (Perform configuration management), SUP.3 (Perform quality assurance), SUP.7 (Perform audits), SUP.8 (Perform problem resolution), MAN.4 (Manage subcontractor) and ORG.2 (Define the process). The average capability level of CUS.1 (Acquire software) is less than 1. However, the sample size in this case is too small for this value to be considered significant. From these results, we can imagine that the nine processes have a possibility close to ISO 9001. However, for a more general conclusion, a comparative study clause-by-clause of ISO/IEC 15504 and ISO 9001 is required. For the non-ISO 9001 certification organizations, ENG.6 (Integrate and test system) only attained average capability level greater than or equal to 2.

11.3. Difference in the Capability Levels of ISO 9001 Certified and Non-ISO 9001 Certified Organizations

In this section, we try to answer the question 'Is there a significant difference in the SPICE capability levels achieved by the processes of ISO 9001 certified organizations and those of non-ISO 9001 certified organizations?'

Figure 8 shows the percentage distribution of capability levels of process instances in the ISO 9001 certified OUs and non-ISO 9001 certified OUs. Comparison of the two pie charts renders the impression that the ISO 9001 certified OUs have greater capability than the non-ISO 9001 certified OUs. For example, 3% of the processes associated with ISO

[10]When the sample size is small, the generic method of assuming a normal distribution cannot be used to construct the confidence interval (Montgomery *et al*. 1998). Instead, a non-parametric statistical approach called the bootstrap method (Kenett and Zacks 1998, Efron and Tibsshirani 1993) can be utilized to compute the confidence interval. The bootstrap method does not depend on a specific distribution function. It samples n times from the original observation with replacement and then computes a sample mean. This process is repeated M times, where M is a large number. The distribution of the M sample means is called the empirical reference distribution. From the distribution, the lower and upper limits of the confidence interval have been determined, with 2.5% and 97.5% percentiles, respectively. This bootstrap method should not be confused with the Bootstrap method for process assessment (Kuvaja 1999).

Copyright © 2001 John Wiley & Sons, Ltd.

Table 17. Capability level of SPICE processes of ISO 9001 certified and non-ISO 9001 certified organizations

Process	ISO 9001 certified organizations (group 1) (group 1)		Non-ISO 9001 certified organizations (group 2) (group 2)		$\bar{x}_1 - \bar{x}_2$	One-sided exact
	Number of PIs	Average ($\bar{x}_1$)	Number of PIs	Average ($\bar{x}_2$)		p-value
CUS.1	2	0.50	3	1.00	−0.50	0.400
CUS.2	15	1.87	16	1.50	0.37	0.058
CUS.3	8	1.63	14	0.93	0.70	**0.017**
CUS.4	3	1.00	10	0.90	0.10	0.604
CUS.5	3	1.33	16	0.81	0.52	0.159
ENG.1	13	2.00	4	1.00	1.00	0.134
ENG.2	26	1.77	30	1.40	0.37	0.143
ENG.3	30	1.90	15	1.53	0.37	0.181
ENG.4	21	1.76	11	0.91	0.85	**0.015**
ENG.5	20	1.80	16	0.88	0.93	**0.002**
ENG.6	9	2.00	5	2.20	−0.20	0.481
ENG.7	9	2.00	15	1.53	0.47	0.134
SUP.1	16	1.75	17	1.00	0.75	**0.018**
SUP.2	22	2.09	24	0.79	1.30	**0.000**
SUP.3	7	2.00	11	0.73	1.27	**0.015**
SUP.4	12	1.42	4	0.25	1.17	0.074
SUP.5	8	1.75	4	0.75	1.00	0.176
SUP.6	8	1.38	11	0.91	0.47	0.154
SUP.7	5	2.20	5	0.40	1.80	0.060
SUP.8	15	2.13	8	1.25	0.88	**0.025**
MAN.1	27	1.59	36	1.11	0.48	**0.005**
MAN.2	16	1.00	9	0.33	0.67	0.118
MAN.3	22	1.14	10	0.60	0.54	0.060
MAN.4	4	2.25	2	0.50	1.75	0.200
ORG.1	5	1.60	9	0.67	0.93	0.119
ORG.2	7	2.29	6	0.17	2.12	**0.001**
ORG.3	6	1.00	5	0.20	0.80	0.056
ORG.4	5	1.60	15	0.87	0.73	**0.007**
ORG.5	6	1.00	10	0.80	0.20	0.419

The processes with p-value shown in bold denote the existence of the difference in the capability level distribution of the two groups at the $\alpha = 0.05$ level of significance.

9001 certified OUs are at level 4 while none of the processes associated with the non-ISO 9001 certified OUs is above level 3.

Table 17 shows the number of process instances (PIs) rated and the average capability level achieved for each of the 29 SPICE processes rated both for the ISO 9001 certified and the non-ISO 9001 certified organizations. The average capability level of the ISO 9001 certified organizations is greater than that of the non-ISO 9001 certified organizations for all processes, apart from CUS.1 and ENG.6. Different samples will produce different values for the difference. Thus, we need a statistical test to determine the existence of the true differences in capability level between two groups (the ISO 9001 certified and the non-ISO 9001 certified organizations).

Although the assumption of an interval scale for capability measurement would allow the use of a parametric test for determining the existence of true differences in capability level between the two groups, a non-parametric test suitable for non-normal, small, unbalanced or skew data was employed in this study[11]. This analysis of the difference between the two groups used the capability levels of the 350 process instances from ISO 9001 certified organizations and the 341 process instances from non-ISO 9001 certified organizations.

[11] A popular method of testing for the existence of a true difference is to use a hypothesis test for the difference in means of two independent normal populations. However, our data set is characterized as being small and unbalanced (or skewed), as well as being non-normal. Thus, we used a non-parametric test called the permutation test (Good 1993; StatXact 1998) that accommodates the characteristics of our data set. Most non-parametric tests can be used on data from an ordinal scale (Conte *et al*. 1986).

Copyright © 2001 John Wiley & Sons, Ltd.

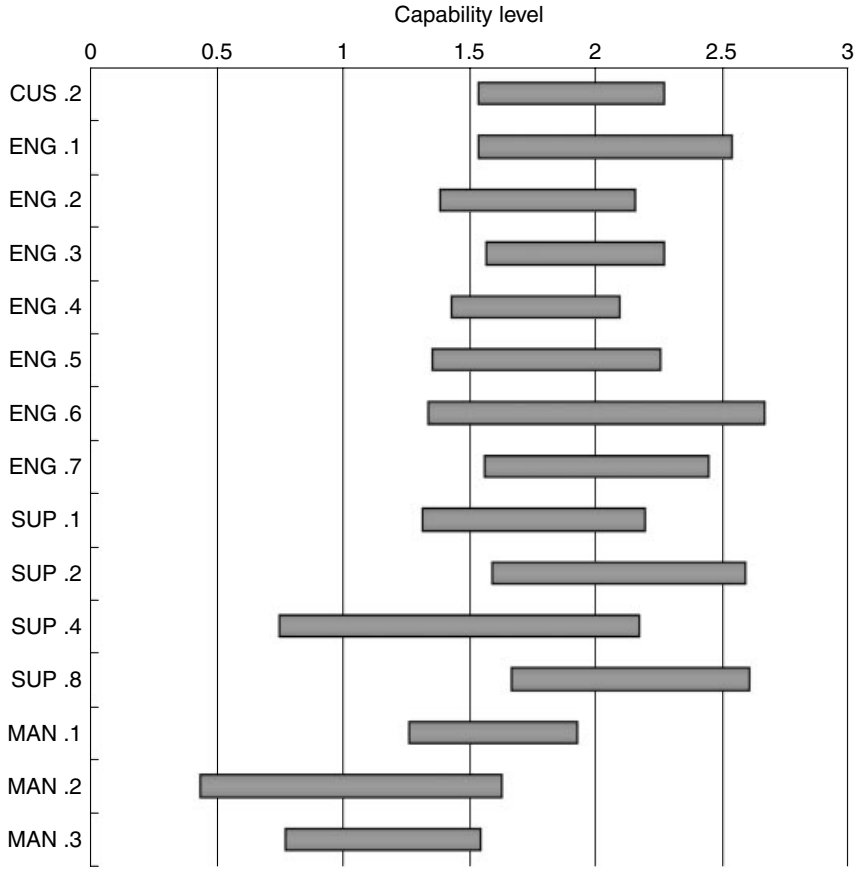

Figure 7. Confidence interval (95%) of the capability levels of ISO 9001 certified organizations (processes with sample size ≥9)

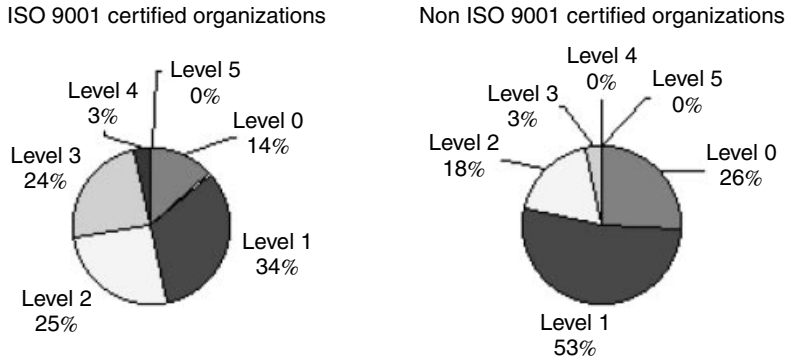

Figure 8. Distribution of capability levels in ISO 9001 certified and non-ISO 9001 certified organizations

The analysis was conducted to determine the same response (capability level) distribution of the two groups. As shown in Table 17, for ten processes CUS.3, ENG.4, ENG.5, SUP.1, SUP.2, SUP.3, SUP.8, MAN.1, ORG.2 and ORG.4, we can say, with a confidence level of 95%, that the capability level of the ISO 9001 certified organizations is greater than that of the non-ISO 9001 certified organizations. For the remaining processes, we cannot say that one group (those associated with ISO 9001 certified

Copyright © 2001 John Wiley & Sons, Ltd.

organizations or those associated with non-ISO 9001 certified organizations) has a significantly different capability distribution from the other.

In addition, we also have compared the response distribution in the capability levels achieved according to the IT staff size. In all SPICE processes except ENG.5 and SUP.2, the average capability level for organizations with a large IT staff is greater than for organizations with a small IT staff (table is omitted). This result seems to suggest that larger companies have more formally defined processes. However, for the sample size of the process instances for the Phase 2 SPICE Trails it is only for ENG.7 and ORG.3 that the capability level distribution of organizations with a large IT staff is statistically greater than that for organizations with a small IT staff.

11.4. Final Remarks

In interpreting the results obtained by comparing SPICE and ISO 9001, the similarities between the requirements of the two standards should be borne in mind. In this sense the results are perhaps not too surprising, though it is encouraging to have empirical evidence to suggest that SPICE and ISO 9001 applied to software, produce broadly similar results.

ISO 9001 (2000) uses a process-based approach and includes some changes to the requirements of the IPPN version of ISO 9001. In addition, ISO 9000-3 (1997) (*Application of ISO 9001:2000 to Software*) is under revision based on software processes, especially on ISO/IEC 12207:1995 (*Software Life Cycle Processes*). Therefore, ISO/IEC 15504 and the latest version of ISO 9001 are much more consistent than before. It is expected that this consistency between the two standards would make ISO 9001 certified OUs achieve higher capability levels in terms of ISO/IEC 15504 than presented here.

One limitation of this study should be made clear in interpreting our results. This limitation is not unique to our study, but is characteristic of most comparison studies. It is worth explaining here. Suppose that an experimenter is interested in investigating the effect of a specific binary factor (i.e. ISO 9001 certification and non-ISO 9001 certification) to determine whether this factor has a significant effect on an observed quantity. In retrospective studies of this sort designed to 'look into the past' (Agresti 1996), observed data is obtained through the analysis of historical data concerning the system or process (Montgomery *et al*. 1998). Thus, retrospective studies of this sort are limited in that they cannot consider the possibilities that non-ISO 9001 certified companies did without the certification process because it was not necessary for their business, albeit having satisfied all the clauses of ISO 9001. This limitation may be solved by performing a randomized experiment that is a more appropriate way of investigating the effect of such factors. However, sometimes such randomized experiments are not possible due to cost, ethics or for legal reasons. Cost is a major barrier preventing randomization studies in the SPICE trials data collection.

12. RESEARCH LIMITATIONS

Trial analyses have a number of limitations that should be made clear in the interpretation of our results. These limitations are not unique to our studies, but are characteristic to most of the process assessment literature. However, it is worth explaining here.

SPICE Trials is the first to use multiple imputation to fill in the missing value in software engineering empirical studies. However, some analyses in two sections (assessor effort model and the comparison between ISO/IEC 15504 and ISO 9001) did not carry out multiple imputation.

The validity study only covered four software processes. Further predictive validity studies are required for the remaining processes. When spreading the 691 process instances into 29 software processes, sample size may not be enough to give statistical power. Then, Phase 3 of SPICE Trials should cover this validity analysis with a higher priority. The sample size issue is not a unique problem in the validity study.

One limitation of evaluating the exemplar model (Part 5) is that only one research method is employed, namely a questionnaire survey. Ideally, one would conduct multiple evaluative studies and then draw conclusions about the strengths and weaknesses of the assessment model. However, it should be recognized that the SPICE Trials are an on-going program of research that employs multiple methods (El-Emam and Goldenson 1995). This is a form of triangulation whereby we 'investigate a phenomenon using a combination of empirical research methods. The intention is that a combination of techniques complement each other' (Wood *et al*. 1999).

Copyright © 2001 John Wiley & Sons, Ltd.

A vexing issue when performing empirical trials is the extent of their generalizability. In the context of the SPICE Trials, specifically, different models and methods can meet the requirements of ISO/IEC 15504, but we do not evaluate all possible models and methods. Therefore, the question is to what extent can our findings be generalized to all assessments that meet the requirements of ISO/IEC 15504.

Based on public statements that have been made thus far, it is expected that some of the more popular assessment models and methods will meet the requirements for compatibility with the reference model (ISO/IEC 15504: Part 2). For example, Bootstrap version 3.0 claims compliance with ISO/IEC 15504, and the CMMI product suite is expected to be 'consistent and compatible'. The assessments from which we obtained our data are also considered to be compliant. The extent to which our results, obtained from a subset of compliant assessments, can be generalized to all compliant assessments is an empirical question and can be investigated through replications of our study. These issues may be a new research direction in SPICE Trials.

13. FINAL SUMMARY

The Phase 2 SPICE Trials was a unique empirical exercise in software engineering standardization. Analyzing the Trials' data gave fruitful results. This section summarizes lessons in the learned as well as future research directions concerning the strengths and weaknesses of the emerging standard ISO/IEC 15504. The weaknesses should be overcome in the next revision of ISO/IEC 15504, while the strengths should be carefully monitored to check the effectiveness in evolving the standard.

The results in Section 6 show that the emerging standard ISO/IEC 15504 has been used in consistency with the two contexts for performing SPA, i.e. identifying strengths, weaknesses and risks inherent in the processes (capability determination) and providing the drivers for prioritizing improvements to processes (SPI).

The capability measure of ISO/IEC 15504 consists of two dimensions: Process Implementation (up to level 3) and Quantitative Process Implementation (levels 4 and 5). The Cronbach alpha coefficient of each capability dimension shows a high enough internal consistency to be useful in practice. In addition, the four-category scales of measuring process attributes, F, L, P, N, cannot be improved in terms of internal consistency by reducing it to a three- or a two-category scale. However, the value of Cronbach alpha coefficient should be monitored for its value in evolving the standard.

Interrater agreement showed high reliability in the assessment of software processes. The four-category scales of measuring process attributes F, L, P, N, cannot be improved in terms of reliability by reducing it to a three- or a two-category scale. This implies that the four-category scale is suitable to use in practice.

The predictive validity study shows that the verisimilitude of the predictive validity premise for small organizations was supported by weak evidence. This may be an indicant that the process capability measure is not appropriate for small organizations, or that the capabilities stipulated in ISO/IEC 15504 do not necessarily improve project performance in small organizations. This issue should be reinvestigated in future studies and then appropriate action should be taken. However, predictive validity for each of the four processes shows that capability level is associated with performance measures.

Nearly all of the respondents used the exemplar assessment model (Part 5) as a source of indicators (95.5%). Approximately 82% of the respondents have used Part 5 intensively. This implies that Part 5 was intensively used in assessments. In general, lead assessors found Part 5 both useful and easy to use. Furthermore, they were satisfied with the level of detail of the exemplar model (87.2%). However, a minority expressed some concern that they could have produced accurate judgments with less detailed evidence. Thus, the next revision of Part 5 can consider reduction of the size of the document. For the capability dimension almost all of the assessors were confident about their understanding up to level 3 attributes. However, the confidence level dropped for levels 4 and 5 in the perceived consistency and repeatability of their judgments. Thus, the next revision should take this into account.

In addition, in evaluation of the exemplar model (Part 5), some weaknesses in the scale were found. Lead assessors had experienced difficulties in making the distinction between the L and P (43.68), between the F and L (24.56%), and between the P and N (21%), albeit the Cronbach alpha coefficient of each capability dimension and Kappa values

Copyright © 2001 John Wiley & Sons, Ltd.

are high enough to be used in practice. Thus, the ISO/IEC 15504 standard, especially the exemplar model (Part 5), is recommended to be revised according to the results of Part 5 evaluation.

The studies of assessor effort show that consensus (interrater agreement) is one of the most important factors to reduce cost in assessment. The consensus problem and the extent to which assessors reach high interrater agreement is partially related to issues discussed in the exemplar assessment model. Improvement of the exemplar assessment model can expect to decrease the concerns at the high cost of assessments. In addition, since high interrater agreement can reduce the consolidation effort in assessments of only 'Organizational' type processes, it is suggested that WG10 formally or informally introduces a classification of processes into 'Organizational' and 'Project' type processes. Then, assessors can pay most attention to the ratings of 'Organization' type processes in order to reduce the cost of the assessment, and also that future research should focus on improving the reliability of rating this type of process.

The capability level for each of the 29 software processes in ISO 9001 certified OUs is higher than that in non-ISO 9001 certified OUs except two processes with small sample size. The (sample) average capability level for each SPICE process lies between 1 and 2.3 in ISO 9001 certified OUs. Since ISO 9001 (2000) uses a process-based approach and includes some changes to the requirements of ISO 9001 (1997), ISO/IEC 15504 and the latest version of ISO 9001 are much more consistent than before. Therefore, it is expected that this consistency between the two standards would tend to make ISO 9001 certified OUs achieve higher capability levels in terms of ISO/IEC 15504 than presented here.

This is not the end of empirical studies for SPICE Trials. It is necessary to perform the replication of our studies during the Phase 3 SPICE Trials in order to have confidence in the findings here.

APPENDIX A: COMPONENTS OF ISO/IEC TR 15504

Figure A1 shows the nine components of the ISO/IEC 15504 (PDTR and TR versions), and indicates the relationships between them.

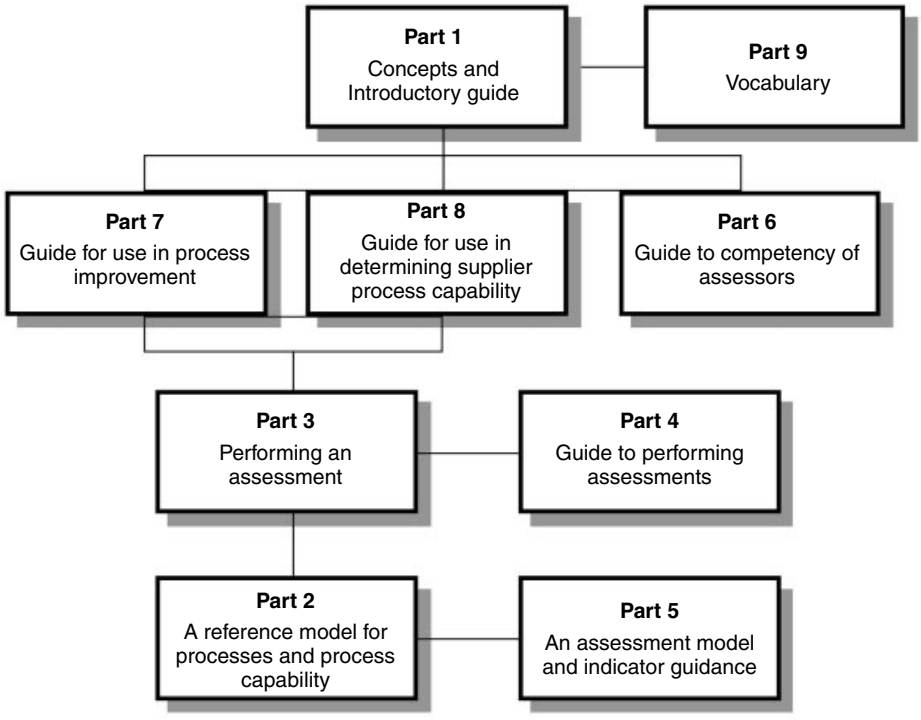

Figure A1. Components of ISO/IEC 15504 documents

Copyright © 2001 John Wiley & Sons, Ltd.

Part 1 is an entry point into ISO/IEC 15504. It describes how the parts of the document suite fit together, and provides guidance for their selection and use. It explains the requirements contained within the standard and their applicability to the performance of an assessment.

Part 2 defines a two-dimensional reference model for describing the outcomes of process assessment. The reference model defines a set of processes, defined in terms of their purpose, and a framework for evaluating the capability of the processes through assessment of process attributes structured into capability levels. Requirements for establishing the compatibility of different assessment models with the reference model are defined.

Part 3 defines the requirements for performing an assessment in such a way that the outcomes will be repeatable, reliable and consistent.

Part 4 provides guidance on performing software process assessments and interpreting the requirements of Part 3 for different assessment contexts. The guidance covers the selection and use of: a compatible assessment model; a supportive method for assessment; an appropriate assessment instrument or tool. This guidance is generic enough to be applicable across all organizations, and also for performing assessments using a variety of different methods and techniques, and supported by a range of tools.

Part 5 provides an exemplar model for performing process assessments that is based upon and directly compatible with the reference model in Part 2. The assessment model extends the reference model through the inclusion of a comprehensive set of indicators of process performance and capability.

Part 6 describes the competence, education, training and experience of assessors that are relevant to conducting process assessments. It describes mechanisms that may be used to demonstrate competence and to validate education, training, and experience.

Part 7 describes how to define the inputs and to use the results of an assessment for the purposes of process improvement. The guide includes examples of the application of process improvement in a variety of situations.

Part 8 describes how to define the inputs and to use the results of an assessment for the purpose of process capability determination. It addresses process capability determination in both straightforward situations and in more complex situations involving, for example, future capability. The guidance on conducting process capability determination is applicable either for use within an organization to determine its own capability, or by an acquirer to determine the capability of a (potential) supplier.

Part 9 is a consolidated vocabulary of all terms specifically defined for the purposes of ISO/IEC 15504.

APPENDIX B: ISO/IEC 15504 PROCESSES AND PROCESS ATTRIBUTES

Table B1. The processes in Process Dimension

Customer-Supplier process category (CUS)
Processes that have direct impact on the customer, support development and software transition to the customer, and provides the correct operation and use software of products and/or services. Its processes are:

- CUS.1: Acquire software
- CUS.2: Manage customer needs
- CUS.3: Supply software
- CUS.4: Operate software
- CUS.5: Provide customer service

Engineering process category (ENG)
Processes that directly specify, implement, or maintain the software product, its relation to the system and its customer documentation. Its processes are:

- ENG.1: Develop system requirements and design
- ENG.2: Develop software requirements
- ENG.3: Develop software design
- ENG.4: Implement software design
- ENG.5: Integrate and test software
- ENG.6: Integrate and test system
- ENG.7: Maintain system and software

Support process category (SUP)
Processes that may be employed by any of the other processes (including other supporting processes) at various points in the software life cycle. Its processes are:

- SUP.1: Develop documentation
- SUP.2: Perform configuration management
- SUP.3: Perform quality assurance
- SUP.4: Perform work product verification

Copyright © 2001 John Wiley & Sons, Ltd.

SUP.5: Perform work product validation
SUP.6: Perform joint review
SUP.7: Perform audits
SUP.8: Perform problem resolution

Management process category (MAN)
Processes which contain generic practices that may be used by those who manage any type of project or process within a software lifecycle. Its processes are:

MAN.1: Manage the project
MAN.2: Manage quality
MAN.3: Manage risks
MAN.4: Manage subcontractors

Organization process category (ORG)
Processes that establish business goals of the organization and develop processes, products and resource assets, which, when used by the projects in the organization, will help the organization achieve its business goals. Its processes are:

ORG.1: Engineer the business
ORG.2: Define the process
ORG.3: Improve the process
ORG.4: Provide skilled human resources
ORG.5: Provide software engineering infrastructure

Table B2. Process attributes for each capability level (ISO/IEC 15504: Part 5)

Capability level	Process attribute (PA)
Level 0 Incomplete	There is a general failure to attain the purpose of the process. There are little or no easily identifiable work products or outputs of the process. Thus, there are no process attributes.
Level 1 Performed process	The purpose of the process is generally achieved. The achievement may not be rigorously planned and tracked. There are identifiable work products for the process, and these testify to the achievement of the purpose. *PA 1.1, Process performance attribute*: The extent to which the process achieves the process outcomes by transforming identifiable input work products to produce identifiable output work products.
Level 2 Managed process	The process delivers work products according to specified procedures and is planned and tracked. Work products conform to specified standards and requirements. *PA 2.1, Performance management attribute*: The extent to which the performance of the process is managed to produce work products that meet the defined objectives. *PA 2.2, Work product management attribute*: The extent to which the performance of the process is managed to produce work products that are appropriately documented, controlled and verified.
Level 3 Established process	The defined process is performed and managed based upon good software engineering principles. Individual implementations of the process use approved, tailored versions of standard, documented processes to achieve the process outcomes. The resources necessary to establish the process definition are also in place. *PA 3.1, Process definition attribute*: The extent to which the performance of the process uses a process definition based upon a standard process to achieve the process outcomes. *PA 3.2, Process resource attribute*: The extent to which the process draws upon suitable resources (for example, human resources and process infrastructure) that is appropriately allocated to deploy the defined process.
Level 4 Predictable process	The defined process is performed consistently in practice within control limits to achieve its process goals. Detailed measures of performance are collected and analyzed. This leads to a quantitative understanding of process capability and an improved ability to predict and manage performance. Performance is quantitatively

Copyright © 2001 John Wiley & Sons, Ltd.

	managed. The quality of work products is quantitatively known. *PA 4.1, Measurement attribute*: The extent to which product and process goals and measures are used to ensure that performance of the process supports the achievement of the defined goals in support of the relevant business goals. *PA 4.2, Process control attribute*: The extent to which the process is controlled through the collection, analysis, and use of product and process measures to correct, where necessary, the performance of the process to achieve the defined product and process goals.
Level 5 Optimizing process	Process performance is optimized to meet current and future business needs, and the process achieves repeatability in meeting its defined business goals. Performance of quantitative process effectiveness and efficiency goals (targets) for performance are established, based on the business goals of the organization. Continuous process monitoring against these goals is enabled by obtaining quantitative feedback and improvement is achieved by analysis of the results. *PA 5.1, Process change attribute*: The extent to which changes to the definition, management and performance of the process are controlled to achieve the relevant business goals of the organization. *PA 5.2: Continuous improvement attribute*: The extent to which changes to the process are identified and implemented to ensure continuous improvement in the fulfillment of the relevant business goals of the organization.

ACKNOWLEDGEMENTS

The authors are members of the SPICE Trials team, and wish to acknowledge the contributions of other past and present members of the trials team, in particular those of Iñigo Garro, Peter Krauth, Bob Smith, Kyungwhan Lee, Angela Tuffley and Alastair Walker. We would also wish to thank all of the LTCs, RTCs, assessors and sponsors of assessments who have made a special effort to ensure that the required data is collected, and who have promoted the SPICE Trials in their regions. In particular, the authors wish to thank Alec Dorling (WG10 Convener) for his consistent support of the work of the SPICE trials. The research of Ho-Won Jung was supported by a Korea University Grant (2001).

The Software Engineering Institute (SEI) is a Federally Funded Research and Development Center sponsored by the U.S. Department of Defense and operated by Carnegie Mellon University.

REFERENCES

Agresti A. 1996. *An Introduction to Categorical Data Analysis*. Wiley: New York.

Aiken L, West S. 1991. *Multiple Regression: Testing and Interpreting Interactions*. Sage University Paper Series on Quantitative Applications in Social Sciences. Sage: Newbury Park.

Belsley DA, Kuh E, Welsch RE. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley: New York.

Briand L, El-Emam K, Moraska S. 1996. On the application of measurement theory in software engineering. *Empirical Software Engineering: An International Journal* **1**(1): 61–88.

Carmines E, Zeller R. 1979. *Reliability and Validity Assessment*. Sage University Paper Series on Quantitative Applications in Social Sciences. Sage: Newbury Park.

Coallier F. 1995. TRILLIUM: A model for the assessment of telecom product development and support capability. *Software Process Newsletter* **2**: 3–8.

Coallier F, Mayrand J, Lague B. 1999. Risk management in software product procurement. In *Elements of Software Process Assessment and Improvement*, El-Emam K, Madhavji NH (eds). IEEE CS Press: California.

Cohen J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **XX**(1): 37–46.

Cohen J. 1968. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**(4): 213–220.

Copyright © 2001 John Wiley & Sons, Ltd.

Conte SD, Dunsmore HE, Shen VY. 1986. *Software Engineering Metrics and Models*. The Benjamin/Cummings Co., Inc.: California.

Cronbach L. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* **16**(3): 297–334.

Dunaway D, Masters S. 1996. CMM-based Appraisal for Internal Process Improvement (CBA IPI): Method Description. Technical Report CMU/SEI-96-TR-007, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Dunaway D, Goldenson D, Monarch I, White D. 1998. How well is CBA IPI working? User feedback. *Proceedings of the SEPG Conference*, Chicago.

Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Chapman and Hall: New York.

El-Emam K. 1998. The internal consistency of the ISO/IEC 15504 software process capability scale. *Proceedings of the 5th International Symposium on Software Metrics* 72–81.

El-Emam K. 1999. Benchmarking Kappa: interrater agreement in software process assessments. *Empirical Software Engineering: An International Journal* **4**(2): 113–133.

El-Emam K, Birk A. 2000a. Validating the ISO/IEC measure of software development process capability. *Journal of Systems and Software* **51**(2): 119–149.

El-Emam K, Birk A. 2000b. Validating the ISO/IEC measure of software requirement analysis process capability. *IEEE Transactions on Software Engineering* **26**(6): 541–566.

El-Emam K, Garro I. 2000. Estimating the extent of standards use: The case of ISO/IEC 15504. *Journal of Systems and Software* **53**(2): 137–143.

El-Emam K, Goldenson D. 1995. SPICE: An empiricist's perspective. *Proceedings of the Second IEEE International Software Engineering Standards Symposium* 84–97.

El-Emam K, Goldenson D. 2000. An empirical review of software process assessments. *Advances in Computers* **53**: 319–423.

El-Emam K, Jung H-W. 2001. An evaluation of the ISO/IEC 15504 assessment model. *Journal of Systems and Software* **49**: 23–41.

El-Emam K, Madhavji NH. 1995. The reliability of measuring organizational maturity. *Software Process–Improvement and Practice* **1**(1): 3–25.

El-Emam K, Marshall P. 1998. Interrater agreement in assessment ratings. In *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*, El-Emam K, Drouin JN, Melo W (eds). IEEE CS Press: California.

El-Emam K, Goldenson D, Briand L, Marshall P. 1996a. Interrater agreement in SPICE based assessments: Some preliminary results. *Proceedings of the Fourth International Conference on the Software Process* 149–156.

El-Emam K, Briand L, Smith R. 1996b. Assessor agreement in rating SPICE processes. *Software Process: Improvement and Practice* **2**(4): 291–306.

El-Emam K, Quintin S, Madhavji H. 1996c. User participation in the requirements engineering process: An empirical study. *Requirement Engineering* **1**: 4–26.

El-Emam K, Smith R, Fusaro P. 1997. Modelling the reliability of SPICE based assessments. *Proceedings of the Third International Symposium on Software Engineering Standards* 69–82.

El-Emam K, Drouin J-N, Melo W (eds). 1998a. *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press: California.

El-Emam K, Simon J-M, Rousseau S, Jacquet E. 1998b. Cost implications of interrater agreement for software process assessment. *Proceedings of the 5th International Symposium on Software Metrics* 38–51.

Fayad ME, Laitinen M. 1997. Process assessment considered wasteful. *Communications of the ACM* **40**(11): 125–128.

Fenton N, Page S. 1993. Towards the evaluation of software engineering standards. *Proceedings of the Software Engineering Standards Symposium* 100–107.

Fenton N, Littlewood B, Page S. 1993. Evaluating software engineering standards and methods. In *Software Engineering: A European Perspective*, Thayer R, McGettrick A (eds). IEEE CS Press: California.

Fusaro P, El-Emam K, Smith B. 1997. Evaluating the interrater agreement of process capability ratings. *Proceedings of the Fourth International Software Metrics Symposium* 2–11.

Fusaro P, El-Emam K, Smith B. 1998. The internal consistencies of the 1987 SEI maturity questionnaire and the SPICE capability dimension. *Empirical Software Engineering: An International Journal* **3**(2): 179–201.

Gardner P. 1975. Scales and statistics. *Review of Educational Research* **45**(1): 43–57.

Goldenson DR, Herbsleb JD. 1995. After the appraisal: a systematic survey of process improvement, its benefits, and factors that influence success. Technical Report CMU/SEI-95-TR-009, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Copyright © 2001 John Wiley & Sons, Ltd.

Good P. 1993. *Permutation Tests, A Practical Guide to Resampling Methods for Testing Hypotheses*. Springer-Verlag: New York.

Hailey V. 1998. A comparison of ISO 9001 and the SPICE framework. In *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*, El-Emam K, Drouin JN, Melo W (eds). IEEE CS Press: California.

Humphrey W, Curtis B. 1991. Comments on a 'a critical look'. *IEEE Software* **8**(4): 42–46.

Humphrey W, Kitson D, Gale J. 1991. A comparison of US and Japanese software process maturity. *Proceedings of the 13th International Conference on Software Engineering* 38–49.

Hunter R. 1998. SPICE trials assessment profile. *IEEE Software Process Newsletter* **12**: 12–18.

ISO 9001. 2000. Quality systems – Model for quality assurance in design, development, production, installation and servicing. ISO, Geneva, Switzerland.

ISO 9000-3. 1997. Quality management and quality assurance standards – Part 3: Guidelines for the application of ISO 9001:1994 to the development, supply, installation and maintenance of computer software. ISO, Geneva, Switzerland.

ISO/IEC N944R. June 1992. The Need and Requirements for a Software Process Assessment Standard. Study Report N944R, Issue 2.0, ISO/IEC JTC1/SC7.

ISO/IEC WG10/N017R. June 1993. Requirements Specification for a Software Process Assessment Standard. Document WG10/N017R Issue 1.00, ISO/IEC JTC1/SC7/WG10.

ISO/IEC JTC1 Directive. 1999. Procedures for the technical work of ISO/IEC JTC1. ISO/IEC JTC1. ISO, Geneva, Switzerland http://www.jtc1.org/directives/toc.htm.

ISO/IEC PDTR 15504. 1996. Information Technology – Software Process Assessment: Part 1–Part 9.

ISO/IEC TR2 15504. 1998. Information Technology – Software Process Assessment: Part 1–Part 9 (Part 5 was published in 1999). ISO, Geneva, Switzerland.

ISO/IEC 12207. 1995. Software Life Cycle Processes. ISO, Geneva, Switzerland.

Jaccard J, Turrisi R, Wan C. 1990. *Interaction Effects in Multiple Regression*. Sage University Paper Series on Quantitative Applications in Social Sciences. Sage: Newbury Park.

Jung H-W. 2001a. Rating the process attribute utilizing AHP in SPICE-based process assessments. *Software Process–Improvement and Practice* **6**(2): 111–122.

Jung H-W. 2001b. Internal and external reliabilities in ISO/IEC TR 15504 assessments. Working paper, Korea University.

Jung H-W, Hunter R. 2001a. The relationship between ISO/IEC 15504 process capability levels, ISO 9001 certification and organization size: an empirical study. *Journal of Systems and Software* **59**: 43–55.

Jung H-W, Hunter R. 2001b. Factors impacting assessor effort in software process assessments: An empirical study. Submitted for publication.

Jung H-W, Hunter R. 2001c. An evaluation of SPICE rating scale at point of internal consistency of capability measure. Submitted for publication.

Kenett RS, Zacks S. 1998. *Modern Industrial Statistics: Design and Control of Quality and Reliability*. Duxbury Press: California, Chapter 7.

Kim J, Mueller C. 1978. *Factor Analysis: Statistical Methods and Practical Issues*. Sage University Paper Series on Quantitative Applications in Social Sciences. Sage: Newbury Park.

Kitchenham BA. 1998. A procedure for analyzing unbalanced datasets. *IEEE Transactions on Software Engineering* **24**(4): 278–301.

Kitson D. 1997. An emerging international standard for software process assessment. *Proceedings of the Third IEEE International Software Engineering Standards Symposium* 83–90.

Kuvaja P. 1999. BOOTSTRAP 3.0 – A SPICE conformant software process assessment methodology. *Software Quality Journal* **8**(1): 7–19.

Maclennan F, Ostrolenk G. 1995. The SPICE Trials: validating the framework. *Proceedings of the 2nd International Symposium* 109–118.

McGarry F, Burke S, Decker B. 1998. Measuring the impacts individual process maturity attributes have on software projects. *Proceedings of the 5th International Software Metrics Symposium* 52–60.

McIver J, Carmines E. 1981. *Unidimensional Scaling*. Sage University Paper Series on Quantitative Applications in Social Sciences. Sage: Newbury Park.

Montgomery DC, Runger GC, Hubele NF. 1998. *Engineering Statistics*. Wiley: New York.

Copyright © 2001 John Wiley & Sons, Ltd.

Narula S, Wellington J. 1977. Prediction, linear regression, and minimum sum of relative errors. *Technometrics* **19**: 185–190.

Nunnally JC, Bernstein IH. 1994. *Psychometric Theory*. McGraw-Hill: New York.

Paulk MC. 1995. How ISO 9001 compares with the CMM. *IEEE Software* **12**(1): 74–83.

Paulk M. 1998. Top-Level Standards Map: ISO 12207, ISO 15504 (Jan 1998 TR) Software CMM v1.1 and v2 Draft C (available at http://www.sei.cmu.edu/pub/cmm/Misc/standards-map.pdf). Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Paulk MC, Webber CV, Garcia SM, Chrissis M, Bush M. 1993. Key Practices of the Capability Maturity Model, version 1.1. Technical Report CMU/SEI-93-TR-25, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Pfleeger S-L, Fenton N, Page S. 1994. Evaluating software engineering standards. *Computer* **27**(9): 71–79.

Rout TP, Tuffley A, Cahill B, Hodgen B. 2000. The rapid assessment of software process capability. *Proceedings of SPICE 2000* 47–55.

Rubin D. 1987. *Multiple Imputation for Nonresponse in Surveys*. Wiley: New York.

Rugg D. 1993. Using a capability evaluation to select a contractor. *IEEE Software* **10**(4): 36–45.

Saiedian H, Kuzara R. 1995. SEI capability model's impact on contractors. *Computer* **28**(1): 16–26.

Sanders M (ed.). 1998. *The SPIRE Handbook: Better, Faster, Cheaper Software Development in Small Organizations*. The SPIRE Project Team: ESSI Project 23873.

SEI. 1999. 'A' Specification for the CMMI Product Suite, version 1.4 (available at http://www.sei.cmu.edu/cmmi/org-docs/aspec1.4.html#scope). Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

SEI. 2000. CMMI for Systems Engineering/Software Engineering, Version 1.02, Continuous Representation. Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality. *Biometrika* **52**: 591–612.

Simon J-M, El-Emam K, Jacquet S-R, Babet F. 1997. The reliability of ISO/IEC PDTR 15504 assessments. *International Software Engineering Research Network*, Technical Report ISERN-97-28.

Smith B, El-Emam K. 1996. Transitioning to Phase 2 of the SPICE Trials. *Proceedings of SPICE'96* 45–55.

SPICE Trials. 1998. SPICE Phase 2 Trials Interim Report, Version 1.00. ISO/IEC JTC1/SC7/WG10.

SPICE Trials. 1999. SPICE Phase 2 Trials Final Report, Vol. 1. ISO/IEC JTC1/SC7/WG10.

StatXact-4 for Windows. 1998. Software for exact nonparametric inference. Cytel Software Co.: Cambridge, MA.

Stevens S. 1951. Mathematics, measurement, and psychophysics. In *Handbook of Experimental Psychology*, Stevens S (ed.). Wiley: New York.

TickIT. 1999. A Guide to Software Quality Management System Construction and Certification Using EN29001. Issue 4.0. UK (TickIT web site http://www.tickit.org).

Velleman P, Wilkinson L. 1993. Normal, ordinal, interval, and ratio typologies are misleading. *American Statistician* **47**: 65–72.

White H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**(4): 817–837.

Wood M, Daly J, Miller J, Roper M. 1999. Multi-method research: an investigation of object-oriented technology. *Journal of Systems and Software* **48**(1): 13–26.

Copyright © 2001 John Wiley & Sons, Ltd.