The Reliability of ISO/IEC PDTR 15504 Assessments

Jean-Martin SIMON

A.Q.T.
19, place de la Ferrandière
69003 Lyon
France
jms.aqt@wanadoo.fr

Khaled El EMAM

Fraunhofer Institute for
Experimental Software
Engineering
Sauerwiesen 6
D-67661 Kaiserslautern
Germany
elemam@iese.fhg.de

Sonia ROUSSEAU(1)
Eric JACQUET(2)

(1)SANOFI Recherche
sonia.rousseau@tls1.elfsanofi.fr
(2)SANOFI Pharma
eric.jacquet@tls1.elfsanofi.fr
9, rue du Président S. Allende
94256 GENTILLY
France

Frederic BABEY

AFNOR
Unité Conseil
Tour Eurpe
92049 Paris La Défense
Cedex – France
101513.2013@compuserve.com

International Software Engineering Research Network Technical Report ISERN-97-28

The Reliability of ISO/IEC PDTR 15504 Assessments

Jean-Martin SIMON

A.Q.T.
19, place de la Ferrandière
69003 Lyon
France
jms.aqt@wanadoo.fr

Khaled El EMAM

Fraunhofer Institute for
Experimental Software
Engineering
Sauerwiesen 6
D-67661 Kaiserslautern
Germany
elemam@iese.fhg.de

Sonia ROUSSEAU[1]
Eric JACQUET[2]

(1)SANOFI Recherche
sonia.rousseau@tls1.elfsanofi.fr
(2)SANOFI Pharma
eric.jacquet@tls1.elfsanofi.fr
9, rue du Président S. Allende
94256 GENTILLY
France

Frederic BABEY

AFNOR
Unité Conseil
Tour Eurpe
92049 Paris La Défense
Cedex – France
101513.2013@compuserve.com

Abstract

During phase two of the SPICE trials, the Proposed Draft Technical Report version of ISO/IEC 15504 is being empirically evaluated. This document set is intended to become an international standard for Software Process Assessment. One thread of evaluations being conducted during these trials is the extent of reliability of assessments based on ISO/IEC PDTR 15504. In this paper we present the first evaluation of the reliability of assessments based on the PDTR version of the emerging international standard. In particular, we evaluate the interrater agreement of assessments. Our results indicate that interrater agreement is considerably high, both for individual ratings at the capability attribute level, and for the aggregated capability levels. In general, these results are consistent with those obtained using the previous version of the Software Process Assessment document set (known as SPICE version 1.0), where capability ratings were also found to have generally high interrater agreement. Furthermore, it was found that the current 4-point scale cannot be improved substantially by reducing it to a 3-point or to a 2-point scale.

1. Introduction

The international SPICE (Software Process Improvement and Capability dEtermination) Project has developed a set of documents describing a model for software process assessment. These documents, known as SPICE version 1.00, were handed over to the ISO/IEC JTC1/SC7 Working Group 10 to evolve them to an international standard. Under the auspices of ISO/IEC, the documents are known by their number 15504. The 15504 documents have to go through a series of ballots by national bodies before they become an International Standard. Subsequent to each ballot, the documents may be changed to address the ballot comments. The most recent balloting stages for 15504 are as follows:

- A *Proposed Draft Technical Report* (PDTR) ballot
- A *Draft Technical Report* (DTR) ballot

Following a successfull DTR ballot, the 15504 documents will become a Technical Report Type 2. This is a designation given to a standard under trial. A TR-2 is expected to be revised within two to three years after its publication, with the intention of making it a full International Standard. A more detailed review of the standardization process for 15504 may be found in [8].

Since the beginning of the effort to develop an international standard for software process assessment, the importance of empirical evaluation of the evolving document set was recognized. This recognition is manifested through the SPICE Trials, which are conducted by the SPICE Project [16]. The first phase of the trials empirically evaluated the SPICE version 1.00 documents, and was completed in calendar year 1995. The second phase of the trials is now underway, and is expected to terminate in the Summer of 1998. This second phase is empicially evaluating the ISO/IEC *PDTR* 15504 document set.

One of the issues studied in the SPICE trials is the reliability of assessments [3]. In general, reliability is concerned with the extent of random measurement error in the assessment scores. There are different types of reliability that can be evaluated. For example, one type is the internal consistency of instruments (see [3][4][14]). This type of reliability accounts for ambiguity and inconsistency amongst indicators or subsets of indicators in an assessment instrument as sources of error. In addition, in the context of the first phase of the SPICE trials, a survey of assessor perceptions of the repeatability of assessments was recently conducted [6].

Interrater agreement is another type of reliability. It is concerned with the extent of agreement in the ratings given by independent assessors to the same software engineering practices. As with many other process assessment methods in existence today (e.g., TRILLIUM-based assessments and the CBA-IPI developed at the SEI), those based on 15504 rely on the judgement of experienced assessors in assigning ratings to software engineering practices. This means that there is an element of subjectivity in their ratings. Ideally, if different assessors satisfy the requirements of the 15504 framework and are presented with the same evidence, they will produce exactly the same ratings (i.e., there will be perfect agreement amongst independent assessors). In practice, however, the subjectivity in ratings will make it most unlikely that there is perfect agreement. The extent to which interrater agreement is imperfect is an empirical question.

High interrater agreement is desirable to give credibility to assessment results, for example, in the context of using assessment scores in contract award decisions. If agreement is low, then this would indicate that the scores are too dependent on the inidividuals who have conducted the assessments. In addition, higher interrater agreement is expected to be associated with lower cost assessments since a consensus-building stage of the assessment method amongst the assessors would consume less time.

During the first phase of the SPICE trials, a number of interrater agreement studies have been conducted [5][7][9][10]. The general conclusion from these studies was that considerable variation in interrater agreement was witnessed, and so models were developed to explain this variation (as in [7]).

The most relevant previous study in the current context is that reported in [13], where elements of the capability dimension were the unit of analysis (as opposed to process instances or processes being the unit of analysis). That study found that interrater agreement is generally high. In this paper we present the first *evaluation* of the interrater agreement of process capability ratings done according to the ISO/IEC PDTR 15504 document set. This evaluation was conducted within the second phase of the SPICE trials.

Briefly, our results indicate that the capability ratings at each of the first three levels of the ISO/IEC PDTR 15504 capability dimension are highly reliable, and that the computed capability levels assigned to these processes are also highly reliable. Furthermore, we found that the current 4-poit scale cannot be substantially improved by combining categories to form 3 or 2 point scales. These results are encouraging for current and potential users since they indicate that assessments using the emerging International Standard maintain high reliability levels after the evolution to the PDTR version.

The next section of the paper provides an overview of the ISO/IEC PDTR 15504 practices rating scheme used during this study. Section 3 presents the research method that was followed for data collection and for evaluating interrater agreement. In section 4 we present the interrater agreement analysis results. We conclude the paper in section 5 with a summary and directions for future work.

2. Overview of ISO/IEC PDTR 15504

2.1 The ISO/IEC PDTR 15504 Document Set

The ISO/IEC PDTR 15504 document set is comprised of nine parts. Figure 1 shows the nine parts of the document set and indicates the relationships between them (see [8] for further details). The most important parts for the current paper are Part 2 and Part 5.

Part 2: A Reference Model for Processes and Process Capability 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 the assessment of process attributes, structured into capability levels. Requirements for establishing the compatibility of different assessment models with the reference model are defined. This part is a normative part of the standard. (Part 2 is described in Chapter 4.)

Part 5: An Assessment Model and Indicator Guidance provides an exemplar model for performing process assessments that is based upon, and is 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.

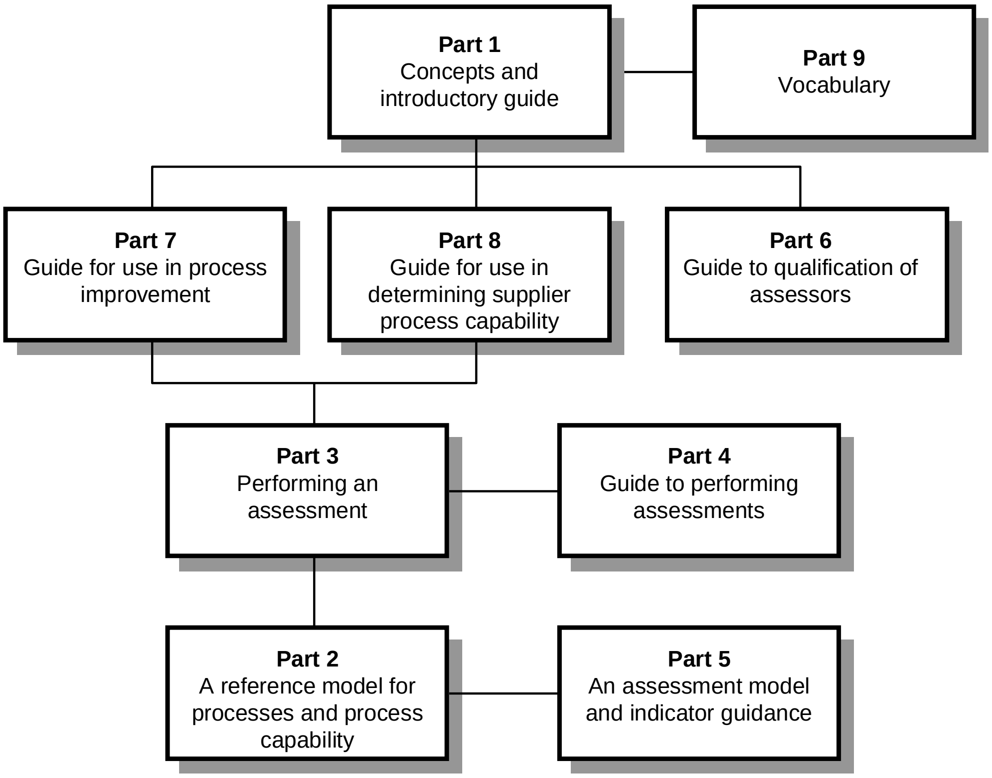
Part 1
Concepts and introductory guide
Part 9
Vocabulary
Part 7
Guide for use in process improvement
Part 8
Guide for use in determining supplier process capability
Part 6
Guide to qualification of assessors
Part 3
Performing an assessment
Part 4
Guide to performing assessments
Part 2
A reference model for processes and process capability
Part 5
An assessment model and indicator guidance

Figure 1: The nine parts of the document set.

2.2 The Capability Rating Scheme in ISO/IEC PDTR 15504

As alluded to above, the ISO/IEC PDTR 15504 architecture is two dimensional. Each dimension represents a different perspective on software process management. The first is the process dimension, and the second is the capability dimension.

The process dimension is divided up into five process categories. Within each category is a set of processes. Each process is characterized by a process purpose. Satisfying the purpose statement of a process represents the first step in building process capability (capabiliy Level 1). The process categories are summarized in Table 1, and their asociated processes are summarized in Table 2.

The capability dimension consists of six capability levels. Within levels 1 to 5 there exists one or two attributes that can be used for evaluating achievement of that level. The levels and their associated attributes are summarized in Table 3. A four-point achievement scale can be used to rate the attributes during an assessment. These are deignated as F, L, P, N, and are summarized in Table 4. It is also possible to convert the F, L. P, N ratings of attributes into a single number that characterizes the capability of a process instance. The scheme for doing so for the first three levels is summarized in Table 5.

Within the context of a ISO/IEC PDTR 15504 assessment, the scope of an assessment is an organizational unit (OU) [8]. This is defined as all or part of an organization with a coherent sphere of activity and a coherent set of business goals. The characteristics that determine the coherent scope of activity - the process context - include the application domain, the size, the criticality, the complexity, and the quality characteristics of its products or services.

Ratings during an assessment are of process instances [8]. A process instance is a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner.

Process Category	Description
Customer-supplier	The *Customer-Supplier* process category consists of 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	The *Engineering* process category consists of processes that directly specify, implement, or maintain a system and software product and its user documentation. In circumstances where the system is composed totally of software, the Engineering process deals only with the construction and maintenance of such software.
Management	The Management process category consists of processes which contain practices of a generic nature which may be used by anyone who manages any sort of project or process within a software life cycle.
Support	The *Support* process category consists of processes which may be employed by any of the other processes (including other supporting processes) at various points in the software life cycle.
Organization	The *Organization* process category consists of processes which establish the business goals of the organization and develop process, product, and resource assets which, when used by the projects in the organization, will help the organization achieve its business goals. Although organizational operations in general have a much broader scope than that of software process, software processes are implemented in a business context, and to be effective, require an appropriate organizational environment.

Table 1: Description of the process categories.

Process Category		Process	
ID	**Title**	**ID**	**Title**
CUS	**Customer Supplier process category**		
		CUS.1	Acquire software
		CUS.2	Manage customer needs
		CUS.3	Supply software
		CUS.4	Operate software
		CUS.5	Provide customer service
ENG	**Engineering process category**		
		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
SUP	**Support process category**		
		SUP.1	Develop documentation
		SUP.2	Perform configuration management
		SUP.3	Perform quality assurance
		SUP.4	Perform work product verification
		SUP.5	Perform work product validation
		SUP.6	Perform joint reviews
		SUP.7	Perform audits
		SUP.8	Perform problem resolution
MAN	**Management process category**		
		MAN.1	Manage the project
		MAN.2	Manage quality
		MAN.3	Manage risks
		MAN.4	Manage subcontractors
ORG	**Organization process category**		
		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 2: The processes and process categories.

ID	Title
Level 0	**Incomplete Process** There is general failure to attain the purpose of the process. There are no easily identifiable work products or outputs of the process.
Level 1	**Performed Process** The purpose of the process is generally achieved. The achievement may not be rigorously planned and tracked. Individuals within the organization recognize that an action should be performed, and there is general agreement that this action is performed as and when required. There are identifiable work products for the process, and these testify to the achievement of the purpose.
1.1	**Process performance attribute**
Level 2	**Managed Process** The process delivers work products of acceptable quality within defined timescales. Performance according to specified procedures is planned and tracked. Work products conform to specified standards and requirements. The primary distinction from the Performed Level is that the performance of the process is planned and managed and progressing towards a defined process.
2.1	**Performance management attribute**
2.2	**Work product management attribute**
Level 3	**Established Process** The process is performed and managed using a defined process based upon good software engineering principles. Individual implementations of the process use approved, tailored versions of standard, documented processes. The resources necessary to establish the process definition are also in place. The primary distinction from the Managed Level is that the process of the Established Level is planned and managed using a standard process.
3.1	**Process definition attribute**
3.2	**Process resource attribute**
Level 4	**Predictable Process** The defined process is performed consistently in practice within defined control limits, to achieve its goals. Detailed measures of performance are collected and analyzed. This leads to a quantitative understanding of process capability and an improved ability to predict performance. Performance is objectively managed. The quality of work products is quantitatively known. The primary distinction from the Established Level is that the defined process is quantitatively understood and controlled.
4.1	**Process measurement attribute**
4.2	**Process control attribute**
Level 5	**Optimizing Process** Performance of the process is optimized to meet current and future business needs, and the process achieves repeatability in meeting its defined business goals. 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. Optimizing a process involves piloting innovative ideas and technologies and changing non-effective processes to meet defined goals or objectives. The primary distinction from the Predictable Level is that the defined process and the standard process undergo continuous refinement and improvement, based on a quantitative understanding of the impact of changes to these processes.
5.1	**Process change attribute**
5.2	**Continuous improvement attribute**

Table 3: Overview of the capability levels and attributes.

Rating & Designation	Description
Not Achieved - N	There is no evidence of achievement of the defined attribute.
Partially Achieved - P	There is some achievement of the defined attribute.
Largely Achieved - L	There is significant achievement of the defined attribute.
Fully Achieved - F	There is full achievement of the defined attribute.

Table 4: The four-point attribute rating scale.

Scale	Process Attributes	Rating
Level 1	Process Performance	Largely or Fully
Level 2	Process Performance	Fully
	Performance Management	Largely or Fully
	Work Product Management	Largely or Fully
Level 3	Process Performance	Fully
	Performance Management	Fully
	Work Product Management	Fully
	Process Definition and Tailoring	Largely or Fully
	Process Resource	Largely or Fully

Table 5: Scheme for determining the capability level rating for the first three levels.

Instructions for Conducting Interrater Agreement Studies

- For each process, divide the assessment team into two groups with at least one person per group.
- The two groups should be selected so that they both meet the minimal assessor competence requirements with respect to training, background, and experience.
- The two groups should use the same evidence (e.g., attend the same interviews, inspect the same documents, etc.), assessment method, and tools.
- The first group examining any physical artifacts should leave them as close as possible (organized/marked/sorted) to the state that the assessees delivered them.
- If evidence is judged to be insufficient, gather more evidence and both groups should inspect the new evidence before making ratings.
- The two groups independently rate the same process instances.
- After the independent ratings, the two groups then meet to reach consensus and harmonize their ratings for the final ratings profile.
- There should be no discussion between the two groups about rating judgment prior to the independent ratings.

Figure 2: Guidelines for conducting interrater agreement studies.

3. Research Method

3.1 Data Collection

For conducting interrater agreement studies, we divide the assessment team into two groups. In the current study, each of these groups had one assessor. Ideally both assessors should be equally competent in making attribute achievement ratings. In practice, both assessors need only meet minimal competence requirements since this is more congruent with the manner in which the 15504 documents would be applied. Each assessor would be provided with the same information (e.g., all would be present in the same interviews and provided with the same documentation to inspect), and then they would perform their ratings independently[1]. Subsequent to the independent ratings, the two assessors would meet to reach a consensus or final assessment team rating. In the context of the SPICE Project, this overall approach is being considered as part of the trials [3]. General guidelines for conducting interrater agreement studies are given in Figure 2. The actual phases of the assessment method that was followed are summarized below.

3.1.1 Preparation Phase

As required by the ISO/IEC PDTR 15504, we defined the assessment *input* at the beginning of the assessment. This consists of:

a) the identity of the sponsor of the assessment and the sponsor's relationship to the organisational unit being assessed,

b) the assessment purpose including alignment with business goals,

c) the assessment scope including:

- ⇒ the processes to be investigated within the organisational unit,
- ⇒ the highest capability level to be investigated,
- ⇒ the organisational unit that deploys these processes,
- ⇒ the process context

d) the assessment constraints which may include:

- ⇒ availability of key resources,
- ⇒ the maximum amount of time to be used for the assessment,
- ⇒ specific processes or OU's to be excluded from the assessment,
- ⇒ the minimum, maximum or specific sample size or coverage that is desired for the assessment,
- ⇒ the ownership of the assessment outputs and any restrictions on their use,
- ⇒ controls on information resulting from a confidentiality agreement.

[1] Under this requirement, one assessor may obtain information that was elicited by the other assessor, which s/he would have not asked for. The alternative to this requirement is that the two assessors interview the same people at different times to make sure that they only obtain the information that they ask for. However, this requirmeent raises the risk that the interviewees "learn" the right answers to give based on the first interview, or that they volunteer information that was asked by the first assessor but not the second. Furthermore, from a practical perspective, intervieweing the same people more than once to ask the same questions would substantially increase the cost of assessments, and thus the cost of conducting the study. It is for this reason that these studies are referred to as "interrater" agreement since, strictly speaking, they consider the reliability of ratings, rather than the reliability of whole assessments. The study of "interassessment" agreement would involve accounting for variations in the information that is collected by two different assessors during an assessment.

e) the identity of the model used within the assessment,

f) the identity of the assessors, including the competent assessor responsible for the assessment,

g) the identity of assesses and support staff with specific responsibilities for the assessment,

h) any additional information to be collected during the assessment to support process improvement or process capability determination.

During the preparation, an important issue is to collect the context of the organisational unit since the result of the assessment is context dependant. Being "context dependant" can best be explained through an example.

In our example, we can consider two organisations, the first is developing a software package with 2000 users on a worldwide basis; the second is a production department which provides a specific MIS application to 20 users who are in the same building. The way those two organisations should organise their Help Desk in order to provide the best "customer service" (CUS.5, see Table 2) is completely different. For example:

a) The first one established a service level agreement with dedicated resources and formal procedures to handle any request and to manage interviews and questionnaires to appraise user satisfaction.

b) The second one mandated its project leader to log any request and to meet on a regular basis the users to appreciate their level of satisfaction.

In the first case, the actions taken are congruent with the complexity and the magnitude of the requirements. However, the same actions seem exaggerated for the second organisation. The assessors therefore have the responsibility to tune their judgement about the capability attributes for the relevant process according to the context.

The context tackles the following parameters :

a) the size of the organisation being assessed;

b) the number of organisational units involved in the assessment;

c) the demographics of the organisational unit,

d) the application domain of the products or services of the organisational unit, the level of organisational participation in performing the assessment (collecting the information, demonstrating conformance);

e) the maturity of the supplier-sponsor relationship (the level of trust between the organisation and sponsor);

f) the needs of the sponsor;

g) the size, criticality and complexity of the products or services,

h) the characteristics of the project for which the processes are evaluated (Process instance).

3.1.2 Data Collection Phase

To conduct the assessment, we used the interview technique based on the assessement model described in Part 5 of ISO/IEC PDTR 15504, plus documents examination.

If necessary, we provide some additional base practices to the model Part 5 for some processes where we deem the Part 5 is too vague. For example, for the CUS.3 Process, we added the following base practices to the CUS.3.7 *Deliver and install software*:

a) CUS.3.7.0 Identify requirements for replication, packaging, storage, handling before delivery

b) CUS.3.7.1 Identify Infrastructure Environment for delivery

c) CUS.3.7.2 Identify training requirements for the client for delivery

d) CUS.3.7.3 Identify duties from the customer or the client for delivery

e) CUS.3.7.4 check delivery before installation

f) CUS.3.7.5 Perform the installation of the software

g) CUS.3.7.6 validate the installation

For all of the processes within the scope of the assessment, only levels 1 to 3 were covered.

3.1.3 Ratings Phase

Each assessor collected his own assessment record during the interview. At the end of the day, each assessor took some time to review his own record and to make the process attributes ratings. Therefore, a specific meeting is dedicated to consolidate the assessment record and to establish a consensus between the 2 assessors when some divergence arises for one or several attribute ratings. This aspect is very important since one of the assessors may have missed or misunderstood some information. In the case that both assessors have missed some information, the sponsor (or the interviewee(s)) is contacted to obtain the missing information.

3.1.4 Debriefing

At the end of the assessment week (the number of days may depend on the number of assessed processes), the 2 assessors present to the interviewees the main results of the assessment. The objectives of this presentation are:

a) to remind them about the concepts of ISO/IEC PDTR 15504

b) to ensure the understanding of the meaning of the attributes by the interviewees,

c) to consolidate with the interviewees the results of the assessment.

During this meeting, the interviewees have the opportunity to “negotiate” the results by, for example, presenting further evidence. At this time, the results are only presented using a graphical approach.

3.1.5 Reporting

We performed the final assessment report where we synthesised the results (weaknesses and strengths) process per process at the OU level. This global analysis is completed with the detailed analysis result for every assessed process for the considered project. This report is therefore sent to the sponsor for approval.

3.2 Description of Organization and Projects

In our study, we used data from two assessments that were conducted in France during the Phase 2 of the SPICE trials. In these assessments, the ISO/IEC PDTR 15504 documents were used. The company where the assessments was conducted is called Sanofi.

The Sanofi company belongs to the ELF Group. Its activities focus on drug research and production. All pharmaceutical molecules must undergo six to twelve long years of development from the moment of their discovery to the time they are given product licence approval. Sanofi R&D has 2,500 employees, in nine units located in six countries (France, UK, Italy, Hungary, Spain and USA). From the research stage on the compound, to international commercialisation, Sanofi R&D controls each phase to test scientifically both the indications for and the effects of the compounds.

The IS (Information Systems) departments interact with all of these activities as a support service. Computerized systems are necessary for several domains: discovery, preclinical studies, clinical investigation, and support. Development methods are either conventional (V model) or prototype based. Software packages are largely used. The architecture is still "mainframe" for some systems, but mostly "Client-Server". IS departments manage the computerized systems life cycle from the initialisation of the system to the retirement. They are used to work closely with users and with the support of the Research Quality Assurance.

Two OUs within this company were assessed. A combination of organizational and project level processes were assessed in each OU. Three projects were assessed in the first OU and two projects in the second OU. The characteristics of these five projects are summarized inTable 6. The processes that were assessed and the number of instances in each are summarized inTable 7.

	X1	**X2**	**X3**	**Y1**	**Y2**
Size of project in terms of effort	3 man-years	2,5 man-year	1 man-years	2 man-years	1 man-years
Programming language,	C, Visual Basic + off-the-shelf software	Third generation language	specific SQL	C, Visual Basic + off-the-shelf software	specific SQL
Development or maintenance projects	maintenance	maintenance	validation	maintenance	maintenance
Application domain	Electronic document management	data processing : collection, processing, visualisation	data base , Client-server	Electronic document management	data base, Client-server

Table 6: Characteristics of assessed projects.

Process	Number of Instances
ORG.1	2
ORG.2	2
ORG.3	2
ORG.4	2
ORG.5	2
CUS.3	5
CUS.4	4
CUS.5	5
ENG.7	4
SUP.1	4
SUP.2	4
MAN.1	4
Total	40 process instances

Table 7: Number of instances of each process assessed.

3.3 Description of Assessors

The same two assessors conducted both assessments. Both assessors met the minimal requirements stipulated in the ISO/IEC PDTR 15504 documents. In terms of experience and background, this is summarized in Table 8.

Both assessors who took part in our study were external. A previous study [9] identified potential systematic biases between an external and an internal assessor (i.e., one assessor would systematically rate higher or lower than the other). Having only external assessors removes the possibility of this particular bias.

	Assessor A	Assessor B
years in the software industry	14	3
years in process assessment and improvement	7 (including software quality improvement)	2
assessment methods & models they have experience with	ISO 9001, SPICE V1, and ISO/IEC PDTR 15504	ISO 9001, Bootstrap, and ISO/IEC PDTR 15504
number of SPICE-based assessments done in the past	6 (approximately 150 process instances)	3 (approximately 90 process instances)
internal vs. external to the organization	external	external

Table 8: Experience and background of assessors.

3.4 Evaluating Interrater Agreement

To evaluate interrater agreement, we can treat the ISO/IEC PDTR 15504 achievement ratings as being on a nominal scale. Cohen [1] defined coefficient Kappa (κ) as an index of agreement that takes into account agreement that could have occured by chance. The value of Kappa is the ratio of observed excess over chance agreement to the maximum possible excess over chance agreement. See [11] for the details of calculating Kappa.

If there is complete agreement, then $\kappa=1$. If observed agreement is greater than chance, then $\kappa>0$. If observed agreement is less than would be expected by chance, then $\kappa<0$. The minimum value of κ depends upon the marginal proportions. However, since we are interested in evaluating agreement, the lower limit of κ is not of interest.

The variance of a sample Kappa has been derived by Fleiss et al. [12]. This would allow testing the null hypothesis that $\kappa=0$ against the alternative hypothesis $\kappa\neq0$. If we use a one-tailed test, then we can test against the alternative hypothesis $\kappa>0$, which is more useful. This means we test whether a value of Kappa bigger than zero as large as the value obtained could have occured by chance.

The standard version of the Kappa coefficient assumes that all disagreements are equally serious. We used a weighted version of Kappa that allows different levels of seriousness to be attached to different levels of disagreement. This has been defined in [2]. The weighted version of Kappa was used in previous studies on the reliability of process assessments [5][13]. We also use the same weighting scheme as applied in previous studies in the SPICE trials [5][13]. This assigns greater seriousness to disagreements on non-adjacent categories on the four-point Achievement scale, and hence esentially treats it as an ordered scale.

3.5 Interpreting Interrater Agreement

After calculating the value of Kappa, the next question is “how do we interpret it?” A commonly used set of guidelines in previous interrater agreement studies (e.g., see [5][13]) are these of Landis and Koch [15].

In addition, we can determine whether the obtained value of Kappa meets a minimal requirement (following the procedure in [11]). The logic for a *minimal* requirement is that it should act as a good discriminator between assessments conducted with a reasonable amount of rigor and precision, and those where there was much misunderstanding and confusion about how to rate practices. It was thus deemed reasonable to require that agreement be at least moderate (i.e., Kappa > 0.4). This minimal requirement on interrater agreement has been used in previous studies in the SPICE trials that evaluate the reliability of process capability ratings [13].

Kappa Statistic	Strength of Agreement
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost Perfect

Figure 3: The interpretation of the values of Kappa.

We evaluate whether interrater agreement using weighted Kappa is greater than moderate agreement for each of the five attributes in levels 1 to 3 of the capability dimension. When performing so many statistical tests, the probability of incorrectly rejecting one of these null hypotheses (Type I error) is approximately 0.4. This means that there is reasonably high probability that at least one significant result would be found by chance alone. We therefore use a Bonferroni adjusted alpha level for hypothesis tests on the 5 attributes in our study (see [17] for an overview of the Bonferroni procedure).

4. Results

The results of evaluating interrater agreement for the five capability attributes are shown in Table 9. As can be seen, ratings on all five attributes have at least moderate agreement at an experiment-wise alpha value of 0.1. These results concur in general with evaluations of interrater agreement of capability ratings for the previous version of the document set (known as SPICE Version 1.0) [13][14].

For the interrater agreement of capability level ratings for each of the processes, the results also indicate statistical significance at an alpha level of 0.1 (see Table 9). Agreement was also found to be consistently higher than “moderate agreement”.

The combination of these results indicates that whether one uses the attribute ratings or the capability ratings, their reliability is higher than moderate agreement. If it is accepted that moderate agreement is a minimal for practical usage, then these results are encouraging for users of ISO/IEC PDTR 15504.

In order to investigate possible sources of disagreement on the 4-point scale, we calculated the weighted Kappa coefficient for the following two cases:

1. 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).

2. 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 10. As can be seen, in most cases the 4-point scale provides the highest Kappa values when compared with the 3- or 2- category scales. The zero value for Attribute 2.2 is due to the data set exhibiting very little variation when reduced to a 2-point scale, and this tends to attenuate the values of Kappa. The conclusion from this table is that the 4-point scale cannot be improved in terms of reliability by reducing it to a 3 or a 2 point scale.

It should be noted that these results have limitations in terms of their generalizability. First, further research is necessary to determine whether similar results would be obtained for a different pair of assessors. While both assessors who took part in this study met the requirements for qualified assessors as stipulated in the ISO/IEC PDTR 15504 documents, further empirical investigation is necessary to ascertain whether *any* assessors that meet these requirments can attain such interrater agreement results. Second, the assessments from which our data were collected were conducted using a particular assessment method. This method is similar to the method used in previous interrater agreement studies [13][14]. However, it remains to be investigated whether the usage of different methods will produce similar results.

Attribute #	Description of Attribute	Weighted Kappa Value	Interpretation
1.1	*Process performance attribute* The extent to which the execution of the process uses a set of practices that are initiated and followed using identifiable input work products to produce identifiable output work products that are adequate to satisfy the purpose of the process.	0.78*	Substantial
2.1	*Performance management attribute* The extent to which the execution of the process is managed to produce work products within stated time and resource requirements.	0.64*	Substantial
2.2	*Work product management attribute* The extent to which the execution of the process is managed to produce work products that are documented and controlled and that meet their functional and non-functional requirements, in line with the work product quality goals of the process.	0.60*	Moderate
3.1	*Process definition attribute* The extent to which the execution of the process uses a process definition based upon a standard process, that enables the process to contribute to the defined business goals of the organization.	0.64*	Substantial
3.2	*Process resource attribute* The extent to which the execution of the process uses suitable skilled human resources and process infrastructure effectively to contribute to the defined business goals of the organization.	0.86*	Almost Perfect
Capability Level	Process capability calculated according to scheme in Table 5.	0.70*	Substantial

Table 9: Interrater agreement evaluation results (* indicates statistical significance).

Attribute #	4-Category	3-Category	2-Category
1.1	0.78	0.59	0.78
2.1	0.64	0.42	0.56
2.2	0.60	0.64	0
3.1	0.64	0.52	0.63
3.2	0.86	0.84	0.79

Table 10: Comparing Achievement scales with different numbers of response categories.

5. Conclusions

In this paper we have presented the method and results of a study to evaluate the interrater agreement of the ISO/IEC PDTR 15504 emerging international standard for software process assessment. The study was based on two assessments conducted in France during the second phase of the SPICE trials. The results of the study indicate that the interrater agreement of these assessments was high, raising confidence in the usage of this version of the 15504 document set for process assessments. In addition, we found that the interrater agreement cannot be improved by reducing the scale to a 3-point nor to a 2-point scale.

Further studies of interrater agreement are planned during the second phase of the SPICE trials. As well as evaluations, we plan to develop models to explain the variation in the reliability of assessments in order to provide guidelines for increasing reliability.

6. Acknowledgements

This study is supported by ELF *Innovation* to promote software process assessment as a new technique for Software Quality Management.

7. References

[1] J. Cohen: “A Coefficient of Agreement for Nominal Scales”. In *Educational and Psychological Measurement*, XX(1):37-46, 1960.

[2] J. Cohen: “Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit”. In *Psychological Bulletin*, 70(4):213-220, October 1968.

[3] K. El Emam and D. R. Goldenson: “SPICE: An Empiricist’s Perspective”. In *Proceedings of the Second IEEE International Software Engineering Standards Symposium*, pages 84-97, Canada, August 1995.

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

[5] K. El Emam, D. R. Goldenson, L. Briand, and P. Marshall: "Interrater Agreement in SPICE Based Assessments: Some Preliminary Results". In *Proceedings of the Fourth International Conference on the Software Process*, pages 149-156, December 1996.

[6] K. El Emam and D. R. Goldenson: "An Empirical Evaluation of the Prospective International SPICE Standard". In *Software Process Improvement and Practice Journal*, 2(2):123-148, 1996.

[7] K. El Emam, R. Smith, and P. Fusaro: “Modelling the Reliability of SPICE Based Assessments”. In *Proceedings of the International Symposium on Software Engineering Standards*, pages 69-82, 1997.

[8] K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capabiliy Determination*. IEEE CS Press, 1997.

[9] K. El Emam, L. Briand, and R. Smith: “Assessor agreement in rating SPICE processes”. In *Software Process Improvement and Practice Journal*, 2(4):291-306, John Wiley, 1997.

[10] K. El Emam and P. Marshall: “Interrater agreement in assessment ratings”. In K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press, 1997.

[11] J. Fleiss: *Statistical Methods for Rates and Proportions*, John Wiley & Sons, 1981.

[12] J. Fleiss, J. Cohen, and B. Everitt: “Large Sample Standard Errors of Kappa and Weighted Kappa”. In *Psychological Bulletin*, 72(5):323-327, 1969.

[13] P. Fusaro, K. El Emam, and B. Smith: “Evaluating the interrater agreement of process capability ratings”. In *Proceedings of the Fourth International Software Metrics Symposium*, pages 2-11, 1997.

[14] P. Fusaro, K. El Emam, and B. Smith: “The Internal Consistencies of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension”. To appear in *Empirical Software Engineering: An International Journal*, Kluwer Academic Publishers, 1997.

[15] J. Landis and G. Koch: “The Measurement of Observer Agreement for Categorical Data”. In *Biometrics*, 33:159-174, March 1977.

[16] F. Maclennan and G. Ostrolenk: “The SPICE Trials: Validating the Framework”. In *Software Process Improvement and Practice Journal*, 1:47-55, 1995.

[17] J. Rice: *Mathematical Statistics and Data Analysis*. Duxbury Press, 1987.