Interrater Agreement in SPICE-Based Assessments: Some Preliminary Results

KHALED EL EMAM

DENNIS R. GOLDENSON

LIONEL BRIAND

PETER MARSHALL

International Software Engineering Research Network technical report ISERN-97-08.

Interrater Agreement in SPICE-Based Assessments: Some Preliminary Results

Khaled El Emam[a,1]
Dennis R. Goldenson[b,2]
Lionel Briand[a]
Peter Marshall[c]

[a]Fraunhofer Institute for Experimental Software Engineering, Germany
[b]Software Engineering Institute, U.S.A.
[c]Defence Research Agency, U.K.

Abstract

The international SPICE Project intends to deliver an ISO standard on software process assessment. This project is unique in software engineering standards in that there is a set of empirical trials, the objectives of which are to evaluate the prospective standard and provide feedback before standardization. One of the enduring issues being evaluated during the trials is the reliability of assessments based on SPICE. One element of reliability is the extent to which different teams assessing the same processes produce similar ratings when presented with the same evidence. In this paper we present some preliminary results from two assessments conducted during the SPICE trials. In each of these assessments two independent teams performed the same ratings. The results indicate that in general there is at least moderate agreement between the two teams in both cases. When we take into account the severity of disagreement then the extent of agreement between the two teams is almost perfect. Also, our results indicated that interrater agreement is not the same for different SPICE processes. The findings reported in this paper provide guidance for future studies of interrater agreement in the SPICE trials and also indicate some potential issues that need to be considered within the prospective standard.

1 Work done by El Emam in the SPICE project has been supported, in part, by the Applied Software Engineering Centre (ASEC) in Montreal.

2 Work done at the SEI is sponsored by the U.S. Department of Defense.

1 Introduction

The international SPICE (Software Process Improvement and Capability dEtermination) Project aims to deliver an ISO standard for software process assessment [13]. As part of this project, there are empirical trials scheduled [3][12]. The empirical trials are divided into three broad phases. The first phase was completed in calendar year 1995. One of the issues studied in this phase was the reliability of assessments based on the SPICE framework [3]. For the developers and users of software process assessments, reliability has been an issue of enduring concern [4][3].

Interrater agreement is one type of reliability (other types include, for example, the internal consistency of instruments [4]). It is concerned with the extent of agreement in the ratings given by independent assessors to the same organizational practices after being presented with the same evidence[3]. As with many other process assessment methods in existence today (e.g., those based on TRILLIUM and Software Capability Evaluations developed at the SEI), those based on SPICE rely on the judgement of experienced assessors in assigning numbers to software engineering practices[4]. This means that there is an element of subjectivity in their

3 This is different from inter-assessment agreement. With inter-assessment agreement one evaluates agreement of ratings from independent assessments where the evidence presented to the assessors may not necessarily be the same. Elevated costs have precluded the conduct of inter-assessment agreement studies thus far.

4 Of course in these assessments the judgements are frequently informed through interviews, document inspections, and also questionnaires.

Process Category	Description
Customer-supplier	processes that directly impact the customer, supporting development and transition of the software to the customer, and provide for its correct operation and use
Engineering	processes that directly specify, implement or maintain a system and software product and its user documentation
Project	processes which establish the project, and co-ordinate and manage its resources to produce a product or provide services which satisfy the customer
Support	processes which enable and support the performance of the other processes on a project
Organization	processes which establish the business goals of the organization and develop process, product and resource assets which will help the organization achieve its business goals

Figure 1: Brief description of the process categories.

ratings. Ideally, if different assessors follow the stipulations of the SPICE 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 individuals who have conducted the assessments.

To our knowledge, there has been no systematic empirical investigations of interrater agreement in the assessment of software processes (however, empirical investigations of the internal consistency of assessment instruments [4][10] and evaluations of assessor perceptions of the repeatability of SPICE-based assessments [5] have been conducted).

In this paper we first present a method for evaluating interrater agreement within the context of a single assessment. Second, through two case studies, we draw some preliminary conclusions about the reliability of assessments based on the SPICE framework.

Briefly, our results indicate that, when taking the severity of disagreement in ratings into account, two independent teams approach almost perfect agreement. When we assume that all disagreements are equally serious, we obtain different results for the different processes that were assessed, but in general the agreements are at least moderate. We discuss these results, provide some guidance for conducting interrater agreement studies based on the lessons learned, and establish an initial baseline to compare future research results.

The next section of the paper provides an overview of the SPICE practices rating scheme that was proposed in the version of the documents used during phase 1 of the trials. Section 3 presents a method for evaluating interrater agreement within the context of a single assessment. In section 4 we present our case studies and the interrater agreement analysis results. We conclude the paper in section 5 with a summary and directions for future work.

2 The Proposed Practices Rating Scheme in SPICE

The SPICE architecture is two dimensional[5]. Each dimension represents a different perspective on software process management. One dimension consists of *processes*. Each process contains a number of *base practices*. A base practice is defined as a software engineering or management activity that addresses the purpose of a particular process. Processes are grouped into *Process Categories*. An example of a process is *Develop System Requirements and Design*. Base practices that belong to this process include: *Specify System Requirements*, *Describe System Architecture*, and *Determine Release Strategy*. An overview of the process categories is given in Figure 1.

The other dimension consists of *generic practices*. A generic practice is an implementation or institutionalization practice that enhances the

5 Since the completion of the first phase of the SPICE trials, elements of the SPICE architecture have been revised and restructured. The basic two dimensional architecture, remains however. In this study, we used the first version of the SPICE documents only.

Capability Level	Description
Level 0 Not Performed	There is general failure to perform the base practices in the process. There are no easily identifiable work products or outputs of the process.
Level 1: Performed-Informally	Base practices of the process are generally performed, but are not rigorously planned and tracked. Performance depends on individual knowledge and effort. There are identifiable work products for the process.
Level 2: Planned-and-Tracked	Performance of the base practices in the process is planned and tracked. Performance according to specified procedures is verified. Work products conform to specified standards and requirements.
Level 3: Well-Defined	Base practices are performed according to a well-defined process using approved, tailored versions of the standard, documented process.
Level 4: Quantitatively-Controlled	Detailed measures of performance are collected and analyzed leading 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.
Level 5: Continuously-Improving	Quantitative process effectiveness and efficiency goals for performance are established, based on the business goals of the organization. Continuous process improvement against these goals is enabled by quantitative feedback.

Figure 2: Brief description of the process categories.

Rating & Designation	Description
Not Adequate - N	The generic practice is either not implemented or does not to any degree satisfy its purpose.
Partially Adequate - P	The implemented generic practice does little to contribute to satisfy the purpose.
Largely Adequate - L	The implemented generic practice largely satisfies its purpose.
Fully Adequate - F	The implemented generic practice fully satisfies its purpose.

Figure 3: Brief description of the rating scheme for the generic practices.

capability to perform a process. Generic practices are grouped into *Common Features*, which in turn are grouped into *Capability Levels*. An example of a Common Feature is *Disciplined Performance*. A generic practice that belongs to this Common Feature stipulates that data on performance of the process must be recorded. An overview of the Capability Levels is given in Figure 2.

Initially each base practice within a process is rated to determine whether the process is actually performed. Once this has been established, each generic practice is rated based on its implementation in the process. This rating utilizes a four-point adequacy scale. The four discrete values are summarized in Figure 3. The four values are also designated as F, L, P, and N.

3 Evaluating Interrater Agreement

In order to evaluate interrater agreement, an assessment must be conducted in a manner that provides the appropriate data. A suitable approach is to divide the assessment team into k groups. It is assumed that each group's assessors are equally competent in making practice adequacy judgements. Ideally, this would be achieved through either random assignment or matching. The assessor(s) in each group 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. For evaluating interrater agreement the k independent ratings would then be compared. The nature of this comparison will be discussed below. Subsequent to the independent ratings, the k groups would meet to reach a consensus or final assessment team rating. In the context of SPICE, this overall approach is being considered as a standard part of the trials [3]. General guidelines for conducting interrater agreement studies are given in Figure 4.

Instructions for Conducting Interrater Agreement Studies

- For each SPICE process, divide the assessment team into two groups with at least one person per group.
- The two groups should be selected so that they are as closely matched as possible 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 SPICE profile.
- There should be no discussion between the two groups about rating judgement prior to consensus building and harmonization[6].

Figure 4: Guidelines for conducting interrater agreement studies.

3.1 A Basic Measure of Interrater Agreement

To simplify the presentation, we assume that k (number of independent ratings) is equal to two[7] . To evaluate interrater agreement[8], we treat the SPICE adequacy ratings as being on a nominal scale. We can then tabulate an assessment's results as shown in Figure 5. In this table P_{ij} is the proportion of ratings classified in cell (i,j), P_{i+} is the total proportion for row i, and P_{+j} is the total proportion for column j:

	Team 2				
Team 1	**F**	**L**	**P**	**N**	**Total**
F	P_{11}	P_{12}	P_{13}	P_{14}	P_{1+}
L	P_{21}	P_{22}	P_{23}	P_{24}	P_{2+}
P	P_{31}	P_{32}	P_{33}	P_{34}	P_{3+}
N	P_{41}	P_{42}	P_{43}	P_{44}	P_{4+}
Total	P_{+1}	P_{+2}	P_{+3}	P_{+4}	1.00

Figure 5: Notation for presenting proportions of ratings in each of the four rating categories by two teams.

$$P_{i+} = \sum_{j=1}^{4} P_{ij}$$

$$P_{+j} = \sum_{i=1}^{4} P_{ij}$$

The most straightforward approach to evaluating agreement is to consider the proportion of ratings upon which the two teams agree:

$$P_O = \sum_{i=1}^{4} P_{ii}$$

However, this value includes agreement that could have occurred by chance. For example, if the two teams employed completely different criteria for assigning their ratings to the same practices (i.e., if the row variable was independent from the column

6 This requirement needs special attention when the assessment method stipulates having multiple consolidation activities throughout an assessment (e.g., at the end of each day in an assessment). Observations that are discussed during such sessions can be judged as organizational strengths or weaknesses, and therefore the ratings of the two teams would no longer be independent. This can be addressed if consolidation is performed independently by the two groups. Then, before the presentation of draft findings to the organization, independent ratings are given followed by consensus building and harmonization of ratings by both teams.

7 This is consistent with the manner in which our case studies were conducted, and therefore makes it easier to understand the case studies and their results. Restricting k to 2 does not result in any loss of generality however. Methods for calculating Kappa for studies where k>2 have been described by Fleiss [6].

8 It should be noted that "agreement" is different from "association". For the ratings from two teams to agree, the ratings must fall in the same adequacy category. For the ratings from two teams to be associated, it is only necessary to be able to predict the adequacy category of one team from the adequacy category of the other team. Thus, strong agreement requires strong association, but strong association can exist without strong agreement. For instance, the ratings can be strongly associated and also show strong disagreement.

variable in Figure 5), then a considerable amount of observed agreement would still be expected by chance.

The extent of agreement that is expected by chance is given by:

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

Cohen [1] has defined coefficient Kappa (κ) as an index of agreement. Kappa takes into account agreement by chance:

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

When there is complete agreement between the two teams, P_O will take on the value of 1. The observed agreement that is in excess of chance agreement is given by $P_O - P_e$. The maximum possible excess over chance agreement is $1 - P_e$. Therefore, κ is the ratio of observed excess over chance agreement to the maximum possible excess over chance agreement.

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.

In addition, the variance of a sample Kappa has been derived by Fleiss et al. [9]. 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.

While its application in software engineering has been limited, the Kappa coefficient has been used most notably by researchers in evaluating the reliability of clinical diagnosis. For example, one study considered the reliability of the diagnosis of multiple sclerosis by neurologists [11], and another considered the diagnosis by psychiatrists of patients into a number of mental disorders, such as depression, neurosis, and schizophrenia [6].

3.2 Interpreting Interrater Agreement

After calculating the value of Kappa, the next question is "how do we interpret it?" There are two general approaches for interpreting such measures. The first is with comparison to previously established baselines. However, given that there are no precedents of interrater agreement studies in software engineering, this approach is not feasible. The second approach is to establish some general benchmarks based on factors such as: what has been learned and accepted in other disciplines, experience within our own discipline, and our intuition. As a body of empirical knowledge is accumulated on software process assessments, we would evolve these benchmarks to take account of what has been learned.

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 6: The interpretation of values of Kappa.

We resort to follow the guidelines developed and accepted within other disciplines. To this end, Landis and Koch [11] have presented a table that is useful and commonly applied for benchmarking the obtained values of Kappa. This is shown in Figure 6. In addition, we can test the hypothesis of whether the obtained value of Kappa meets a minimal requirement (following the procedure in [7]). The logic for a minimal benchmark 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 value was perceived as a good discriminator.

It should be cautioned, however, that the benchmark that we suggest above should only be considered initial. If, after further empirical study, it was found that these benchmarks fail all SPICE processes, pass all of them, or pass ones that intuitively should be failed and vice versa, then the benchmark should be modified to strengthen or weaken the requirement.

3.3 Accounting for Seriousness of Disagreement

This κ coefficient assumes that all disagreements are equally serious. An alternative would be to use weighted Kappa [2] if the relative seriousness of the disagreements can be specified. Weighted κ is given by:

$$\kappa = \frac{P_{O(w)} - P_{e(w)}}{1 - P_{e(w)}}$$

where

$$P_{O(w)} = \sum_{i=1}^{4}\sum_{j=1}^{4} w_{ij}P_{ij}$$

$$P_{e(w)} = \sum_{i=1}^{4}\sum_{j=1}^{4} w_{ij}P_{i+}P_{+j}$$

When $w_{ij}=0$ for all cells off the diagonal (i.e., $i \neq j$), then weighted Kappa becomes identical to unweighted Kappa (because this indicates that all disagreements are equally serious). The weighting scheme that we propose would consider disagreements on adjacent categories on the four-point scale as less severe than disagreements on categories that are two or more categories further apart. Without weighting, the four-point scale can be considered to be at the nominal level. With this weighting scheme we are essentially adding ordinal information to the scale (i.e., adjacent categories are "closer" to each other in terms of measuring adequacy). There are many potential weighting schemes that can be used. There are no precedents in software engineering, and therefore we chose a scheme that has been applied in other disciplines and that reflects our intuitive understanding of the severity of disagreements. A suitable weighting scheme has been proposed by Fleiss and Cohen [8]:

$$w_{ij} = 1 - \frac{(i-j)^2}{(C-1)^2}$$

where C is the number of categories, in this case 4. The weights are given in Figure 7.

	Team 2			
Team 1	**F**	**L**	**P**	**N**
F	1	0.89	0.55	0
L	0.89	1	0.89	0.55
P	0.55	0.89	1	0.89
N	0	0.55	0.89	1

Figure 7: Weights to reflect severity of disagreement.

4 Two Case Studies

We used data from two assessments based on the draft SPICE documents that were conducted in Europe during the first phase of the SPICE trials. A description of these assessments and the results are presented in this section.

4.1 Description of Assessments

The first organizational unit that was assessed had a size of approximately 40 personnel who developed real time embedded systems using Ada. For the interrater agreement study, the ENG.3 process (Develop Software Design) was selected. The purpose of this process is to establish a software design that effectively accommodates the software requirements. Practices within this process are: develop software architectural design, design interfaces at top level, develop detailed design, and establish traceability. The assessment team was divided into two teams. The two teams were chosen to be of equal experience. Both teams attended the interviews and therefore had access to the same information. They then performed their adequacy ratings independently.

For the second study, a different organizational unit was assessed. This organizational unit consisted of 95 staff members. The project assessed was staffed by 12 software engineers developing a real-time embedded system of approximately 10 KLOC written in Ada. The conduct of the assessment was similar to the one above. The data we obtained was from assessing the PRO.5 process (Manage Quality). The purpose of this process is to manage the quality of the project's products and services to ensure that they satisfy the customer. Practices within this process are: establish quality goals, define quality metrics, identify and perform quality activities, assess quality, and take corrective action.

4.2 Overall Interrater Agreement

The presentation of our results consists of two elements. First, the computed value of Kappa. Second, the interpretation of the strength of agreement is given. For all values of Kappa that are

	Study 1 (ENG.3 Process - Develop Software Design)			Study 2 (PRO.5 Process - Manage Quality)		
	Proportion Agreement	**Kappa**	**Interpretation**	**Proportion Agreement**	**Kappa**	**Interpretation**
Overall (4-Category scale)	77%	0.59	Moderate	78%	0.70	Substantial
3-Category Scale	80%	0.63	Substantial	84%	0.76	Substantial
2-Category Scale	97%	0.92	Almost Perfect	91%	0.79	Substantial
Only High Capability Lvls	62%	0.46	Moderate	69%	0.54	Moderate
Only Low Capability Lvls	85%	0.44	Moderate	85%	0.72	Substantial

Figure 8: Results from the interrater agreement analysis.

presented in this section, we tested the hypothesis that their values were greater than zero. The null hypothesis was rejected in all cases at a one tailed alpha level of 0.05.

Initially, values for weighted kappa were calculated. These are 0.92 for study one and 0.88 for study two. These values indicate almost perfect agreement. Therefore, taking the severity of disagreement into account, one can claim that these assessments demonstrated a very high level of agreement between the two teams. This can be explained by the fact that most disagreements were between adjacent categories on the four-point scale.

Overall values of unweighted Kappa were then calculated. As can be seen in Figure 8, overall agreement ranges from “moderate” in study 1 to “substantial” in study 2. These levels of interrater agreement both surpass our minimal requirement of being at least "moderate" (statistical test conducted at a one-tailed alpha level of 0.05). Another point to notice is that the difference in the values of Kappa between the two processes is not small. This indicates that interrater agreement is potentially different for different processes. This may be a function of how well the different processes are described in the SPICE documentation and how well they are understood by the assessors in general.

4.3 Sources of Disagreement

To better understand the sources of disagreement, we calculated Kappa for the two following cases:

1 Combining the two middle categories of the adequacy 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 these combinations are also shown in Figure 8. In both cases, the combination of categories increases the value of Kappa. However, for the data from the first study the two category scale results in a larger increase in agreement than the three category scale. This suggests that there is more confusion in rating practices at the end points of the adequacy scale for that process. Conversely, the difference between the two grouping strategies in study 2 are quite small (0.76 vs. 0.79). This indicates that there is possibly equal confusion of categories at the end points of the scale as at the middle points of the scale (according to the two grouping strategies presented above).

Subsequently, we considered the extent of agreement at high capability levels when compared to low capability levels. From overall ratings summaries from the SPICE trials (see [14]), it was clear that there was a paucity of ratings better than

"Not Adequate" for the Engineering and Project process categories at higher maturity levels. Since there are few instances rated highly at the higher capability levels, then this means that there is little knowledge about higher level generic practices. Consequently greater disagreement in higher capability ratings would be expected.

High capability levels include levels 3 to 5. Low capability levels include levels 1 and 2. As can be seen in Figure 8, for the ENG.3 process there is not much difference in the Kappa values (0.46 vs. 0.44). However, for the PRO.5 process, agreement at the low capability levels is considerably larger than agreement at the high capability levels (0.72 vs. 0.54). This may be because assessors have less understanding of processes for managing quality at higher levels of capability (e.g., quantitative control and improvement of quality management practices). We can at least make the preliminary conclusion that for some processes, ratings at high capability levels may be less reliable than those at the low capability levels. But again, the results do differ for different SPICE processes.

5 Conclusions

In this paper we have presented two case studies of interrater agreement in assessments based on SPICE. These evaluated the extent to which independent assessment teams agree in their ratings after being presented with the same evidence. The results indicate that in general there is moderate to substantial agreement in the ratings by two independent teams. In both cases, the interrater agreement values meet minimal benchmark requirements. If we take the severity of disagreements into account, then the extent of agreement is almost perfect. Perhaps most interestingly, we have found that the extent of agreement differed for the different processes that were studied. This indicates that future interrater agreement studies should consider processes independently and should not attempt to pool the data from different processes.

While these results are only preliminary, they do provide an initial baseline with which to compare the results from future studies of interrater agreement, and they do raise some issues that deserve further investigation in the SPICE trials. General studies of interrater agreement are planned in phase 2 of the SPICE trials. Furthermore, investigations of assessor confusion on the SPICE rating scale and for different capability levels are planned.

Acknowledgements

The authors wish to express their sincere gratitude to Harry Barker, Mac Craigmyle, and John Hamilton for making their assessment data available to the SPICE community. The authors also wish to thank Walcelio Melo for his comments on an earlier version of this paper.

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, 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 and D. R. Goldenson: "An Empirical Evaluation of the Prospective International SPICE Standard". In *Software Process Improvement and Practice Journal*, 2(2), 1996.

[6] J. Fleiss: "Measuring Nominal Scale Agreement Among Many Raters". In *Psychological Bulletin*, 76(5):378-382, 1971.

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

[8] J. Fleiss and J. Cohen: "The Equivalence of Weighted Kappa and the Interclass Correlation Coefficient as Measures of Reliability". In *Educational and Psychological Measurement*, 33:613-619, 1973.

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

[10] W. Humphrey and B. Curtis: "Comments on 'A Critical Look'". In *IEEE Software*, pages 42-46, July 1991.

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

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

[13] T. Rout: “SPICE: A Framework for Software Process Assessment”. In *Software Process Improvement and Practice Journal*, Pilot Issue, pages 57-66, August 1995.

[14] I. Woodman and R. Hunter: "Analysis of Assessment Data from Phase One of the SPICE Trials". In *Software Process Newsletter*, IEEE TCSE, No. 6, pages 5-10, Spring 1996.