Assessor Agreement in Rating SPICE Processes*

Khaled El Emam[a]
Fraunhofer Institute for Experimental Software Engineering
Sauerwiesen 6
D-67661 Kaiserslautern
Germany
elemam@iese.fhg.de

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

Robert Smith
European Software Institute
Parque Tecnologico de Zamudio #204
E-48170 Bilbao
Spain
Bob.Smith@esi.es

Abstract

One of the enduring issues being evaluated during the SPICE trials is the reliability of assessments. One type of reliability is the extent to which different assessors produce similar ratings when assessing the same organization and presented with the same evidence. In this paper we report on a study that was conducted to start answering this question. Data was collected from an assessment of 21 process instances covering 15 processes. In each of these assessments two independent assessors performed the ratings. We found that six of the fifteen processes do not meet our minimal benchmark for interrater agreement. Three of these were due to systematic biases by either an internal or external assessor. Furthermore, for eight processes specific rating scale adjustments were identified that could improve its reliability. The findings reported in this paper provide guidance for assessors using the SPICE framework.

1. Introduction

The international SPICE (Software Process Improvement and Capability dEtermination) Project aims to deliver an ISO standard for software process assessment [21]. As part of this project, there are empirical trials scheduled [3][20]. The empirical trials are divided into three broad phases. The first phase was completed in calendar year 1995. The second phase started in the fourth quarter of 1996. One of the issues studied in the SPICE trials is the reliability of assessments based on the SPICE framework [3]. In general, reliability is concerned with the extent of random measurement error in the assessment scores.

For the developers and users of software process assessments, reliability has been an issue of concern [4][3]. Evaluating the reliability of software process assessments can potentially lead to three types of improvements: (a) improvements in practical decisions made using quantitative assessment results [3], (b) empirical studies investigating the relationship between capability ratings and the effectiveness of projects and organizations are likely to produce consistent results if the reliability of measures of capability are taken into consideration [4], and (c) improvements to assessment models and methods to increase the reliability of assessments [7]. Such improvements can, for example, help develop justifications for investment in assessments (point b) and increase our confidence in assessment results for suppliers and purchasers of software products (points a and c).

There are different types of reliability that can be evaluated. For example, one type is the internal consistency of instruments (see [3][4][13]). 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 SPICE trials, a survey of assessor perceptions of the repeatability of assessments was recently conducted [6].

---

* This appears as International Software Engineering Research Network technical report ISERN-96-09, 1996.

[a] Work done by El Emam in the SPICE project and reported upon in this paper has been supported, in part, by the Applied Software Engineering Centre (ASEC) in Montreal.

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 SPICE 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 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 inidividuals who have conducted the assessments.

To our knowledge, there has been only one published systematic empirical investigation of interrater agreement in the assessment of software processes thus far, and this was performed in the context of the SPICE trials [5]. This initial study evaluated the interrater agreement for two SPICE processes. In this paper we report on a subsequent study to evaluate the interrater agreement for the same two processes as before and for another thirteen different SPICE processes.

In our study we evaluate agreement between two individual assessors who rate the same processes. Evaluating the reliability of individual assessor judgements is of value because the first version of the SPICE documents did not explicitly exclude one person assessments [14]. In fact, the size of assessment teams in phase 1 of the SPICE trials is shown in Figure 1 (see [6] for more details of the SPICE phase 1 trials results). Out of 35 assessments, almost 9% were single person assessments. Furthermore, the new version of the SPICE guidance documents, version 2.0, does explicitly allow an assessment team to consist of only one team member, especially for small assessments [16]. This means that single person assessments are acceptable from a SPICE perspective. The general question being addressed then is whether single assessor ratings are repeatable?

# of persons on assessment team	# of assessments	% of assessments
1	3	8.6%
2	14	40%
3	8	22.8%
4	8	22.8%
6	2	5.7%

Figure 1: Number of assessors in the assessment team in phase 1 of the SPICE trials.

Briefly, our results indicate that six out of the fifteen processes assessed do not meet minimal benchmark requirements for interrater agreement. Three of these were due to systematic biases by either an internal or external assessor. Furthermore, for eight processes specific rating scale adjustments were identified that could improve its reliability. We discuss these results and provide some guidance for conducting assessments based on the SPICE framework.

The next section of the paper provides an overview of the SPICE practices rating scheme that has been proposed in the version of the documents used during this study. Section 3 presents the research method that was followed for data collection and for evaluating interrater agreement within the context of a single assessment. 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. The Proposed Practices Rating Scheme in SPICE

The SPICE architecture is two dimensional[1]. 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 2.

The other dimension consists of *generic practices*. A generic practice is an implementation or institutionalisation practice that enhances the 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 3.

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 4. The four values are also designated as F, L, P, and N.

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 2: Brief description of the Process Categories.

[1] Elements of the SPICE architecture have recently 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 processes.
Level 4: Quantitatively-Controlled	Detailed measures of performance are collected and analysed 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 3: Brief description of the capability levels.

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 4: Description of the rating scheme for generic practices.

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 this 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[2]

Figure 5: Guidelines for conducting interrater agreement studies.

3. Research Method

3.1 Data Collection

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 two groups. It is assumed that each group's assessors are equally competent in making practice adequacy judgements. Ideally, this would be achieved through 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. Subsequent to the independent ratings, the two groups would meet to reach a consensus or final assessment team rating. In the context of SPICE, this overall approach is being considered as part of the trials [3]. General guidelines for conducting interrater agreement studies are given in Figure 5.

In our study, we used data from one assessment that was conducted in the UK during the calendar year 1996. In this assessment, the first version of the SPICE documents were used. The company where the assessment was conducted designs, develops and supplies complete aircraft for the international market. A significant proportion of its business is for export and a substantial proportion of its business is derived from the support and upgrade of existing aircraft fleets. At the time of the assessment the company employed 5000 people overall.

The organizational unit where the assessment took place was the Software and Systems Engineering Department. This department, amongst other tasks, specifies overall aircraft systems, monitors the development for software in the bought-in aircraft systems, and provides software for simulators and avionics integration rigs. The projects that were directly assessed were related to the provision of software for avionics integration rigs. The avionics integration rigs are used to dynamically bench test and prove the integrated aircraft avionics before they are fitted to the aircraft, and to investigate anomalies that have been reported from flight trials. Fifty people work on these projects.

[2] 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 groups 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 groups.

The software that provides the functionality of the rigs has been written in Pascal. It is developed on a cluster of VAX computers and workstations, and targeted to Intel 80x86 based single board computers. PC's programmed in Pascal are used for the operator interface and off-line data preparation and analysis. Since the rigs are used to prove the avionics and to save on flight trials, some of which would be impossible and/or dangerous to do for real, the correct functioning of the rigs is critical. Therefore, there are very high reliability and usability requirements on the software. Program sizes are in the order of 100 KSLOC in Pascal.

Fifteen processes were each assessed by two independent experienced assessors. The processes are described in Figure 6. In total 21 process instances were assessed in this manner. One of these assessors was external to the organization, and the second one was internal. In total there were three external assessors and five internal assessors. On average, each external assessor was involved in assessing 5 processes, and each internal assessor was involved in assessing 3 processes.

For each process instance, the two assessors interviewed a staff member on the process instance being assessed. The questions were shared between the two of them and they both were present when answers were given. Each assessor took his/her own notes and made individual preliminary ratings before a discussion between them where the harmonized ratings were made. All of the internal assessors received five days of training on SPICE-based assessments the week prior to the assessment. This course is a basis for training assessors to participate in phase 2 of the SPICE trials.

3.2 Evaluating Interrater Agreement

To evaluate interrater agreement[3], we treat the SPICE adequacy ratings as being on a nominal scale. Cohen [2] has 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 [5] 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 value of Kappa depends strongly on the marginal distributions (see [1]). This means that the same rating procedure can potentially produce different values of Kappa depending on the proportion of each of the adequacy levels that were rated for a given process. However, Kappa does have the advantage of taking into consideration chance agreement. In addition, when compared to perhaps more intuitive indices of agreement such as percentage agreement, Kappa tends to have lower values than percentage agreement [12]. Therefore, Kappa is more conservative. It is then noted in [12] that *"The tradition in science to accept conservative rather than liberal estimates suggests that percentage agreement is the least desirable [when compared to other reliability estimates, including Kappa]"*. Therefore, we have a strong justification for using the coefficient Kappa over percentage agreement.

[3] 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.

Process	Base Practices
Develop System Requirements and Design (ENG.1)	Specify System Requirements and Design Describe System Architecture Allocate Requirements Determine Release Strategy
Develop Software Requirements (ENG.2)	Determine Software Requirements Analyze Software Requirements Determine Operating Environment Impact Evaluate Requirements with Customer Update Requirements for Next Iteration
Develop Software Design (ENG.3)	Develop Software Architectural Design Design Interfaces at Top Level Develop Detailed Design Establish Traceability
Implement Software Design (ENG.4)	Develop Software Units Develop Unit Verification Procedures Verify the Software Units
Integrate and Test Software (ENG.5)	Determine Regression Test Strategy Build Aggregates of Software Units Develop Tests for Aggregates Test Software Aggregates Develop Tests for Software Test Integrated Software
Integrate and Test System (ENG.6)	Build Aggregates of System Elements Develop Tests for Aggregates Test System Aggregates Develop Tests for System Test Integrated System
Maintain System and Software (ENG.7)	Determine Maintenance Requirements Analyze User Problems and Enhancements Determine Modifications for Next Upgrade Implement and Test Modifications Upgrade User System
Perform Joint Audits and Reviews (CUS.4)	Establish Joint reviews and Audits Prepare for Customer Audits and Reviews Conduct Joint Management reviews Conduct Joint Technical Reviews Support Customer Acceptance Review Perform Joint Process Assessment

Figure 6: Description of the base practices in each of the assessed processes.

Process	Base Practices
Establish project Plan (PRO.2)	Develop Work Breakdown Structure Identify Project Standards Identify Specialized Facilities Determine Reuse Strategy Develop Project Estimates Identify Initial Project Risks Identify Project Measures Establish Project Schedule Establish project Commitments Document Project Plans
Manage Quality (PRO.5)	Establish Quality Goals Define Quality Metrics Identify Quality Activities Perform Quality Activities Assess Quality Take Corrective Action
Manage Resources and Schedule (PRO.7)	Acquire Resources Track progress Conduct management Reviews Conduct Technical Reviews Manage Commitments
Perform Configuration Management (SUP.2)	Establish Configuration Management Library System Identify Configuration Items Maintain Configuration Item Descriptions Manage Change Requests Control Changes Build Product Releases Miantain Configuration Item History Report Configuration Status
Perform Quality Assurance (SUP.3)	Select Project Standards Review Software Engineering Activities Audit Work Products Report Results Handle Deviations
Perform Peer Reviews (SUP.5)	Select Work Products Identify New Standards Establish Completion Criteria Establish Re-review Criteria Distribute Review Materials Conduct Peer Review Document Action Items Track Action Items
Define the Process (ORG.2)	Define Goals Identify Current Activities, Roles & Responsibilities Identify Inputs and Outputs Define Entry and Exit Criteria Define Control Points Identify External Interfaces Identify Internal Interfaces Define Quality Records Define Process Measures Document the Standard Process Establish Policy Establish Performance Expectations Deploy the Process

Figure 6: Description of the base practices in each of the assessed processes (contd.).

The variance of a sample Kappa has been derived by Fleiss et al. [11]. 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.

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 [13], and another considered the diagnosis by psychiatrists of patients into a number of mental disorders, such as depression, neurosis, and schizophrenia [10].

3.3 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 such baselines 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.

We resorted to guidelines developed and accepted within other disciplines. To this end, Landis and Koch [13] have presented a table that is useful and commonly applied for benchmarking the obtained values of Kappa. This is shown in Figure 7. Everitt [8] notes that while this table is arbitrary, it is still potentially useful for interpreting values of Kappa.

In addition, we can test the hypothesis of whether the obtained value of Kappa meets a minimal requirement (following the procedure in [9]). 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). Based on the results reported here and other studies already completed [5], 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 this benchmark fails 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.

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 7: The interpretation of the values of Kappa.

4. Results

In this section we present the overall results and their interpretations. The detailed results are presented in the Appendix. We also discuss threats to the validity of our results.

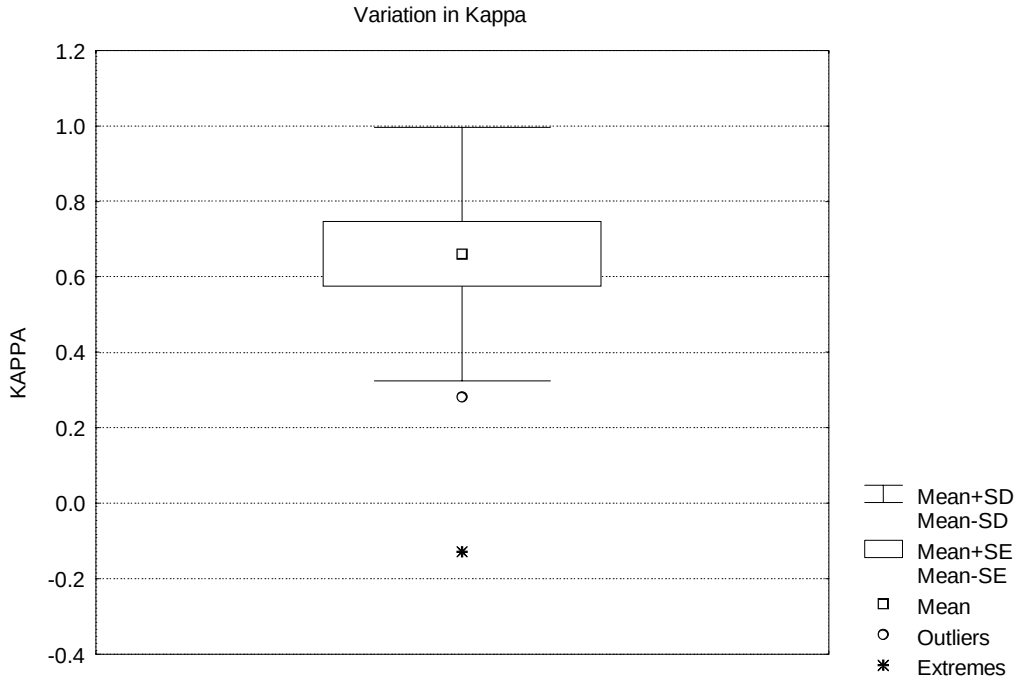

Figure 8: Box and whisker plot showing the variation in the value of Kappa.

4.1 Interrater Agreement

The variation in the extent of interrater agreement was substantial, varying from poor agreement to perfect agreement. This is illustrated in Figure 8. Perfect agreement (i.e., $\kappa=1$) was reached on four out of the fifteen processes (27%). In terms of meeting our minimal requirement of having at least "moderate" agreement, nine out of the fifteen processes passed (60%) and six failed (inferential test was conducted at an alpha level of 0.1). The six processes that failed were ENG.4, ENG.5, ENG.7, SUP.3, SUP.5 and ORG.2.

To better understand the reasons for the high levels of disagreement, we first considered the data distributions. For the ENG.5, even though there was very high agreement at 94%, the ratings were concentrated in one cell. Therefore, the low variation in this data set may be contributing to this process not passing our minimal threshold.

For the remaining processes that failed, we subsequently tested whether there was a systematic difference in the ratings given by the internal assessor versus the external assessor. For that, we used a sign test (see [22]) to determine if either assessor rated higher/lower than the other. The test was conducted at an alpha level of 0.1. For the processes ENG.4 and SUP.5 there were no systematic differences. This means that approximately half of the disagreements were due to the internal assessor rating higher than the external assessor (or the external assessor rating higher than the internal assessor), which is what would be expected by chance. The teams who rated the ENG.4 and SUP.5 process were also completely different, and therefore no assessor specific trends could be hypothesized. Three possible interpretations of this finding are that these two processes are not defined in a manner that is clear enough for objective assessment, that the four-point rating scale is creating confusion in assessing these processes, or that more emphasis should be placed on the method of assessment (e.g., insufficient evidence is collected to make a judgement). Since this is the first study where this type of analysis has been performed for these processes, it is not clear whether the same results would be obtained consistently. However, as a recommendation based on the results thus far, it would be better to be more rigorous in the assessment of these particular processes (e.g., inspection of more evidence and extra corroboration effort). This will ensure that there are sufficient observations to make a reliable rating.

Systematic differences were found for processes ENG.7, SUP.3, and ORG.2. Interestingly, one process' interrater agreement was not better than what would be obtained by chance (process ENG.7), indicating considerable disagreement. For this particular process, the internal assessor was a senior manager (Chief Systems Designer) in the organization and he tended to give much lower ratings than those given by the external assessor. The SUP.3 and ORG.2 processes were rated by the same external assessor. That external assessor tended to give lower ratings than the internal assessor. This possibly indicates a bias either by the external assessor against favouring the organization or by the internal assessors towards favouring the organization. This external assessor did have experience working in a similar environment. For the ratings of the SUP.3 process the internal assessor felt that his higher rating was more accurate because he had inside information some of which was not given by the assessees. In this case it seems that the internal assessor was biased towards the organization. For the ORG.2 process it is not clear whether it was the internal or the external assessor who was systematically biased. The fact that systematic differences were found, however, indicates that there is the potential for biased results if one relies on one type of assessor (e.g., only internal or only external) in an assessment[4]. This further emphasizes the need for consolidation of findings and consensus building during an assessment.

The interrater agreement for two of the same processes (ENG.3 and PRO.5) was evaluated in a previous study [5]. In terms of meeting the minimal requirement, the results concur (although, as expected, the Kappa values are not exactly the same). This further evidence increases our confidence that interrater agreement for these two is sufficient for practical purposes (according to our criteria of at least moderate agreement).

4.2 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 from this analysis also depended on the process being assessed. For the processes that had perfect agreement, there is no confusion, therefore combining categories will have no effect. Of the remaining processes, eight increased their interrater agreement by combining rating categories, which also helps us identify potential confusion amongst categories. Processes ENG.3 and SUP.5 did not benefit from category combination. This may be due to the distribution of the responses (i.e., there was little variation in the data set) however rather than the existence of equal confusion amongst *all* of the categories.

Processes ENG.2, ENG.4, ENG.5, ENG.7, and SUP.2 benefited substantially by reducing the four-point rating scale into a two-point rating scale. Process PRO.2 benefited substantially by reducing the four-point scale into a three-point scale. Finally, processes ENG.6 and ORG.2 benefited from scale reduction, but it seems that there was no particular category combination strategy that would be most useful. Based on an examination of the cell proportions in a 4x4 table for each of ENG.6 and ORG.2, it is evident that there is a similar amount of confuson between all of the adjacent categories of the four-point scale. These latter two would require further investigation to determine whether, for example, an alternative scale altogether may increase agreement or an improved definition of the categories on the scale would be sufficient.

[4] It should be noted that it is not yet known - at least through published empirical research - whether systematic biases will also appear if only internal or only external assessors were used. Therefore, any recommendations on how to act based on this finding can only be tentative.

4.3 Threats to Validity

A number of potential limitations in the form of threats to the validity of our study were considered. These were threats to internal and external validity.

In this study we focused on *evaluating* the reliability of assessments based on the SPICE framework. We implicitly assumed that reliability is only a function of the SPICE documents and architecture (e.g., the clarity of practice definitions, the soundness of the rating scheme, and the applicability of the two-dimensional architecture). Threats to internal validity would question this assumtpion.

One potential threat to internal validity is a *maturation effect*. In this study, a maturation effect would be indicated by a change in interrater agreement (as measured by Kappa) over the course of the assessment. For example, as the assessment progresses, assessors may become more fatigued and pay less attention to observing evidence and in making their ratings. This would tend to decrease the extent of interrater agreement as the assessment progresses. Conversely, assessors may gain knowledge of the organization and the way it implements its practices as time progresses. As more evidence is gathered by assessors they may start to converge in their perceptions about the capability of the organization's processes. This could lead to an increase in interrater agreement as the assessment progresses. If we find a maturation effect then the values of Kappa that we obtained are also a function of when ratings are made during an assessment.

To determine if there was a maturation effect, we conducted a number of post-hoc tests. The assessment ratings were made over a 2.5 day period (the whole assessment was longer since it included an initial meeting with management and a closing session where findings were presented). Evidence on nine processes was inspected and ratings were made in the first 1.5 days of the assessment. These were classified as *early* processes. The remaining six processes were rated in the final day. These were classified as *late* processes. We tested for differences in the values of Kappa between these two groups. We used a two-tailed test at an alpha level of 0.1. The statistic we used was the Mann-Whitney U test [22]. No diferences were found, and hence there is no evidence that the median Kappa values between the two groups differed.

Three different external assessors and five different internal assessors took part in the assessment. The distribution of assessors over time was not uniform, and therefore the maturation effect may be occuring at a different rate for different assessors. For example, one internal assessor took part in assessing only one *late* process, and another took part only in assessing *early* processes. Therefore, for these two assessors there is no maturation effect. An alternative way of measuring progress through the assessment would be the number of processes assessed thus far by the assessors, instead of using time. We calculated the robust Spearman rho coefficient [22] between the number of processes assessed thus far and Kappa. This was done for the internal assessor only, for the external assessor only, and for the sum of the number of proceses assessed thus far for both the internal and external assessor. The rho coefficient was not statistically significant using a two-tailed test at an alpha level of 0.1. Therefore we could not find evidence of a maturation effect. [5]

Another potential threat to validity is a *selection effect*. Where high disagreement was found, differences in capability levels between the internal and external assessor may explain the disagreement. External assessors will tend to have experience with a variety of different organizations and hence more knowledge of different ways of implementing SPICE processes. Also, they would tend to have more experience with assessments. We attempted to counteract this by giving the internal assessors a five day course on SPICE and on process assessments. Internal assessors will tend to have more knowledge of the organization's business, needs, and constraints. However, knowledge of the organization is not considered as a prerequisite in the qualification guidance for SPICE assessors [15]. In terms of general and software education, software training and software experience no discernable differences between the internal and external assessors were

[5] Note that we performed a post-hoc power analysis of these results. Statistical power is the probability that a statistical test will correctly reject the null hypothesis (in this case that the correlation coefficient is zero). We found that the power of the statistical test for these correlations was less than 30% using the tables in [19]. This is a low power level. Therefore, the statistical test used was not powerful enough to detect a maturation effect of the size found in our study. The small sample size is a major contributor to the low power level witnessed here. Similar evaluations using the Pearson correlation, after removal of an outier observation, do not change the general conclusions.

recorded. In terms of the assignment of external assessors to assessing specific processes, this was done randomly. The internal assessors were assigned to projects on which they did not work. Assigning internal assessors randomly would not be advisable as we wanted to ensure that they would not be involved in assessing projects that they had worked on so as not to compromise confidentiality and also to create a climate that encourages the free flow of information between assessees and assessors.

To attain external validity, one could conduct the study with a representative sample from the target population whom we want to generalize to. In the current study such samplng was not performed since all data was collected from one assessment of one organization. Another approach for attaining external validity, but which takes longer, is through replication [18]. When a study is replicated with a different sample and the results are consistent, then they are confirmed despite the differences between the original and replication sample. This lends credence to the generalizability of results. The original SPICE study on the interrater agreement of assessments appeared in [5]. The current study confirmed the original findings as well as presenting new baseline results for the processes not covered in [5]. We are planning further studies of this nature in order to generalize these findings.

5. Conclusions

This paper reported on a field study to evaluate the interrater agreement between independent assessors while rating the same SPICE processes. Some of the results of our study are encouraging for SPICE, while others highlight the need for further empirical investigation of the reliability of process assessments.

In total, we collected data on ratings of fifteen SPICE processes. We found that for four processes the two independent assessors had perfect agreement. Nine processes of the fifteen passed a modest threshold that specifies minimal interrater agreement. Furthermore, the two processes that passed the same threshold in a previous study [5] also passed it in the current study, thus providing for some consistency of results. Based on this evidence, we conclude that the assessment of these nine processes is reliable when we consider different individual raters as the source of error.

Our analysis also revealed different ways for improving interrater agreement. Agreement for some processes improves by combining categories on the 4-point SPICE version 1.0 adequacy scale. In addition, we identified that some assessors may have systematic biases that inflate or undermine their ratings. Our results highlight the need for consolidation and consensus building sessions during an assessment. It is reasonable to assume that the consolidated ratings are more reliable than the ratings of the independent assessors.

Perhaps most importantly from a research perspective, our results make clear that the reliability of process assessments is a serious issue deserving of more concerted empirical investigation. We found that two out of the six ratings of processes did not meet the minimal interrater agreement threshold, were not badly distributed, and did not exhibit any systematic biases by the assessors. This means that the resultant ratings of these processes were substantially affected by the individual making the ratings. While some would like to believe that assessments are sufficiently reliable, our results indicate that this is not always the case. It may be claimed that the above assertion is applicable only to SPICE. However, it should be recalled that SPICE is based largely on existing assessment methods and architectures and the expertise gained in applying them, and therefore a general concern with the reliability of process assessments is warranted. Moreover, to our knowledge, systematic study of interrater agreement of software process assessments outside the scope of the SPICE trials have not yet been conducted, making it more difficult to defend claims supportive of their reliability.

Further research should of course attempt to confirm (or otherwise) the findings presented here. Also, research to date has not covered all of the SPICE processes. Therefore more reliability studies on the processes not covered here are encouraged. In the context of the SPICE trials, larger scale and confirmatory studies of interrater agreement are planned.

Without achieving high reliability levels for capability measures we will not be able to demonstrate the validity of these measures (i.e., that high capability is related to project and organizational

effectiveness). Reliability is a necessary condition for validity. Perhaps most critically then, future research efforts should attempt to investigate the specific factors that have a sizeable impact on the reliability of assessments in order to make recommendations for increasing their reliability.

6. Acknowledgements

The authors wish to thank Alan Davies and Fiona Maclennan for their participation in the data collection. Guiseppe Satriani helped with the data collection, collation of results and producing the assessment report. Finally, our gratitude to the assesment sponsor, the internal assessors and the assessees for their time.

7. References

[1] A. Agresti: An Introduction to Categorical Data Analysis, John Wiley, 1996.

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

[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: “Modeling the Reliability of SPICE Based Assessments”. Submitted for Pulication, 1997.

[8] B. Everitt: *The Analysis of Contingency Tables*, Chapman & Hall, 1992.

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

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

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

[12] D. Hartmann: “Considerations in the Choice of Interobserver Reliability Estimates”. In *Journal of Applied Behavior Analysis*, 10(1):103-116, Spring 1977.

[13] W. Humphrey and B. Curtis: “Comments on ‘A Critical Look’”. In *IEEE Software*, pages 42-46, July 1991.

[14] ISO/IEC JTC!/SC7: “Software Process Assessment Part 4: Guide to Conducting Assessments”. Working Draft 1.0, 1995.

[15] ISO/IEC JTC1/SC7: "Software Process Assessment Part 6: Qualification and Training of Assessors". Working Draft 1.0, 1995.

[16] ISO/IEC JTC1/SC7: “Software Process Assessment Part 4: Guide to Performing Assessments”. Working Draft (revised) 2.0, 1996.

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

[18] R. Lindsay and A. Ehrenberg: “The Design of Replicated Studies”. In *The American Statistician*, 47(3):217-228, 1993.

[19] H. Kraemer and S. Thiemann: *How Many Subjects ? Statistical Power Analysis in Research*, Sage Publications, 1987.

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

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

[22] S. Siegel and J. Castellan: *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, 1988.

8. Appendix: Detailed Results

The following tables contain the detailed results of the analysis that was performed. Where there is an asterisk (*) next to a value of Kappa, that indicates that it is significantly larger than zero at an alpha level of 0.1. We have used a less stringent value of alpha here compared to our previous study in [5] (where an alpha level of 0.05) was used because the number of observations per process in the current study is generally small and therefore a reduction in the power of the statistical test is expected. By increasing the alpha level, we contribute towards increasing the level of statistical power. The tables also show whether the value of Kappa passed our minimal requirement of "moderate" agreement (i.e., Kappa > 0.4). This test was conducted at an alpha level of 0.1 as above.

For processes PRO.7, SUP.3 and SUP.5, three instances were assessed (and hence the larger values of n). For all of the remaining processes, only one instance was assessed.

Develop System Requirements and Design (ENG.1)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	21		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	100%	1*	Almost Perfect
3-Category Scale	100%	1*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect
Develop Software Requirements (ENG.2)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	21		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	95%	0.87*	Almost Perfect
3-Category Scale	95%	0.86*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect
Develop Software Design (ENG.3)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	21		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	90%	0.70*	Substantial
3-Category Scale	90%	0.69*	Substantial
2-Category Scale	95%	0.64*	Substantial
Implement Software Design (ENG.4)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	19		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	57%	0.33*	Fair
3-Category Scale	67%	0.45*	Moderate
2-Category Scale	88%	0.68*	Substantial

Integrate and Test Software (ENG.5)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	18		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	94%	0.65*	Substantial
3-Category Scale	94%	0.64*	Substantial
2-Category Scale	100%	1*	Almost Perfect
Integrate and Test System (ENG.6)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	21		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	75%	0.62*	Substantial
3-Category Scale	85%	0.75*	Substantial
2-Category Scale	89%	0.74*	Substantial
Maintain System and Software (ENG.7)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	13		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	16%	-0.13	Poor
3-Category Scale	62%	0.14	Slight
2-Category Scale	55%	0.20	Slight
Perform Joint Audits and Reviews (CUS.4)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	17		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	100%	1*	Almost Perfect
3-Category Scale	100%	1*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect

Establish Project Plan (PRO.2)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	17		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	88%	0.79*	Substantial
3-Category Scale	94%	0.89*	Almost Perfect
2-Category Scale	94%	0.64*	Substantial
Manage Quality (PRO.5)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	13		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	100%	1*	Almost Perfect
3-Category Scale	100%	1*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect
Manage Resources and Schedule (PRO.7)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	54		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	100%	1*	Almost Perfect
3-Category Scale	100%	1*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect
Perform Configuration Management (SUP.2)			
Passes Minimal Requirement	Yes		
Number of Generic Practices Rated (n)	18		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	93%	0.92*	Almost Perfect
3-Category Scale	93%	0.88*	Almost Perfect
2-Category Scale	100%	1*	Almost Perfect

Perform Quality Assurance (SUP.3)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	54		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	52%	0.28*	Fair
3-Category Scale	63%	0.26*	Fair
2-Category Scale	72%	0.19*	Slight
Perform Peer Reviews (SUP.5)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	51		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	74%	0.50*	Moderate
3-Category Scale	76%	0.52*	Moderate
2-Category Scale	96%	0.48*	Moderate
Define the Process (ORG.2)			
Passes Minimal Requirement	No		
Number of Generic Practices Rated (n)	18		
	Proportion Agreement	**Kappa**	**Interpretation**
Overall (4-Category Scale)	50%	0.37*	Fair
3-Category Scale	72%	0.57*	Moderate
2-Category Scale	71%	0.52*	Moderate