Evaluating the Interrater Agreement of Process Capability Ratings

PIERFRANCESCO FUSARO

KHALED EL EMAM

BOB SMITH

International Software Engineering Research Network Technical Report ISERN-97-20

Evaluating the Interrater Agreement of Process Capability Ratings

Pierfrancesco Fusaro[a]
Khaled El Emam[a]
Bob Smith[b]

[a]Fraunhofer Institute for Experimental Software Engineering, Germany
[b]European Software Institute, Spain

Abstract

The reliability of process assessments has received some study in the recent past, much of it being conducted within the context of the SPICE trials. In this paper we build upon this work by evaluating the reliability of ratings on each of the practices that make up the SPICE capability dimension. The type of reliability that we evaluate is interrater agreement: the agreement amongst independent assessors' capability ratings. Interrater agreement was found to be generally high. We also identify one particular practice that exhibits low agreement in its ratings.

1 Introduction

The objective of the SPICE (Software Process Improvement and Capability dEtermination) Project is to deliver an ISO standard on Software Process Assessment (see [9]). Unique to software engineering standardisation efforts, the SPICE Project includes a set of empirical trials [5][6]. One of the issues receiving substantial empirical study in the trials is the reliability of assessments based on the SPICE framework. Reliability can be defined in general as the extent to which repeated assessments of the same processes will yield the same ratings.

A software process assessment is a measurement procedure that involves expert judgement. It is therefore a subjective measurement procedure. To have confidence in subjective measurement procedures it must be demonstrated that they are reliable. This clearly also applies to process assessments (a review of reliability in the context of process assessments may be found in [5]).

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 individual assessor assessments are being conducted, especially within the context of small organisations. For example, the first version of the SPICE documents did not explicitly exclude one person assessments [15]. In fact, out of 35 assessments that constituted the first phase of the SPICE trials, almost 9% were single person assessments. Furthermore, the subsequent 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]. The Draft Technical Report that has recently been submitted to ISO also allows for single assessor assessments [17]. 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?

The SPICE architecture is two dimensional, with one dimension consisting of the processes that are rated, and the second dimension the capabilities of processes (see Figure 1). Previous research has evaluated the interrater agreement within a process across capabilities (along line A in Figure 1) [7][8]. This approach is useful for evaluating interrater agreement in rating a single process. The current paper evaluates interrater agreement of the individual elements of the capability dimension *across* processes (along line B in Figure 1). This approach allows us to draw more specific conclusions about the reliability of rating each of the elements of SPICE capability (called Generic Practices).

Briefly, our results indicate that most of the Generic Practices of the capability dimension exhibit at least moderate interrater agreement. For one Generic Practice interrater agreement tended to be lower: *Perform Peer Reviews*.

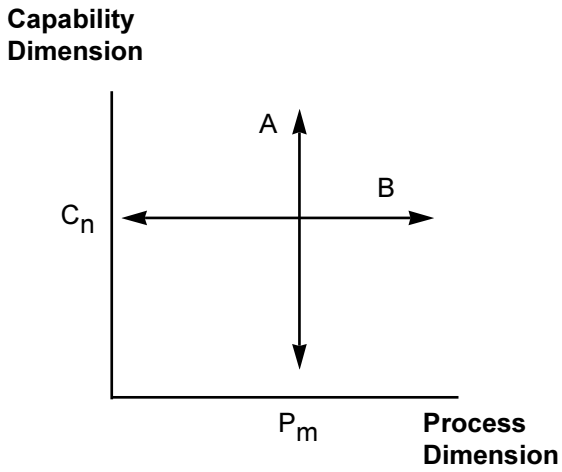

Figure 1: Overview of two approaches for evaluating interrater agreement in SPICE assessments.

The next section of the paper provides an overview of the SPICE practices rating scheme embodied in the version of the documents used during our study, and a description of the variables that we considered. Section 3 presents our research method, including data collection and the analysis method. In section 4 we present the overall results, and discuss their limitations. We conclude the paper in section 5 with a summary and directions for future work.

2 Rating Scheme in SPICE v1.0

In this section we present the rating scheme that is used in version 1.0 of the SPICE documents. This is the version of SPICE that was used during our study.

The SPICE architecture is two dimensional as depicted in Figure 2. 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 3.

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

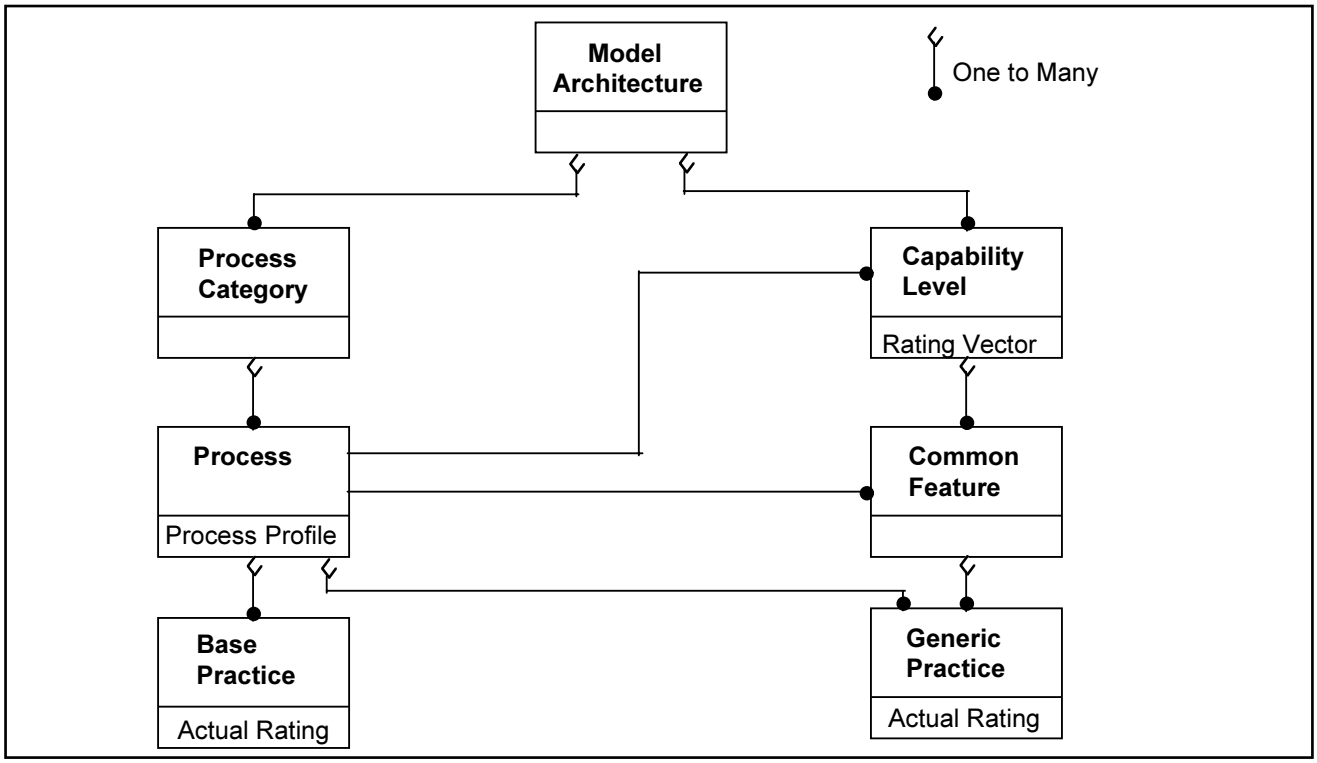

Figure 2: Basic architecture of SPICE Version 1.0 (source [20]).

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
Organisation	processes which establish the business goals of the organisation and develop process, product and resource assets which will help the organisation achieve its business goals

Figure 3: Brief description of the process categories.

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 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 organisation. Continuous process improvement against these goals is enabled by quantitative feedback.

Figure 4: 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 5: Brief description of the rating scheme for the generic practices.

- 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 both meet the minimal SPICE 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 (organised/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 harmonise their ratings for the final SPICE profile.
- There should be no discussion between the two groups about rating judgement prior tothe independent ratings.

Figure 6: Guidelines for conducting interrater agreement studies.

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 4.

Initially each base practice within a process is rated to determine the extent to which the process is actually performed. Once this performance has been established, each subsequent Generic Practice is rated based on its implementation in the process. These ratings utilise a four-point adequacy scale. The four discrete values are summarised in Figure 5. The four values are also designated as F, L, P, and N.

In the current paper we consider the Generic Practices at Levels 1 to 3. These practices are presented in the results section of the paper. In total, there are 18 Generic Practices in these three capability levels.

3 Research Method

In this section we describe the method that was used for collecting the data and for data analysis.

3.1 General Method for Interrater Agreement Studies

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 2 groups. In the current study, each of these groups had one assessor. Ideally, both assessors should be equally competent in making practice adequacy judgements. In practice, both assessors need only meet minimal competence requirements since this is more congruent with the manner in which the SPICE framework would be applied. Each assessor would be provided with the same information (e.g., s/he would be present in the same interviews and provided with the same documentation to inspect), and then s/he would make the ratings independently. Subsequent to the independent ratings, the 2 assessors would meet to reach a consensus or final assessment team rating (this final rating is usually presented to management at the end of the assessment, and serves as the basis for process improvement actions). General guidelines for conducting interrater agreement studies are given in Figure 6.

3.2 Data Collection

The data for this study was obtained from two assessments conducted within the European trials region during 1996. In total, 33 process instances were assessed.

During an assessment ratings are done on *process instances*. A process instance is defined as a *singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner* [20].

The SPICE documents do not define a process for conducting an assessment (usually referred to as the assessment method), although the documents do provide method guidance (see [4]). The method used in this study is therefore only one of many possible methods that can be used in a SPICE-

conformant assessment. One constraint was that the method had to be suitable for providing us with the necessary data by following the guidelines in Figure 6.

The method used for the assessments was as follows. First, there was a half day pre-assessment meeting between the assessors and the organisational unit personnel for introductions and scoping of the assessment. The first half day of the actual assessment consisted of an introduction to SPICE and to the assessment for all of the assessment participants. This is followed by two and a half days of information gathering and process instance ratings. Information was gathered for each process instance to be assessed through interviews and document reviews. Right after, the ratings for that particular process instance were made independently by the two assessors, and then the harmonised ratings were made. The assessment concludes with a half day preparation of the final ratings and a meeting with the assessment sponsor. The assessment is then closed by a 2 hour presentation of the results of the assessment.

Both assessors that took part in our study were experienced and met the minimal qualification requirements for conducting a SPICE-based assessment. In fact, there was little variation in this factor, and therefore it could not explain variation in reliability. This factor was controlled for two reasons. First, we could not justify to the organisations sponsoring the assessment having unqualified or inexperienced assessors. Second, all assessments that are conformant with the SPICE framework are supposed to use qualified assessors; not having qualified assessors is not relevant for understanding SPICE-conformant assessments.

Both assessors who took part in our study were external. A previous study [7] 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.

3.3 Estimation of Interrater Agreement

One approach for evaluating interrater agreement is to treat the SPICE adequacy ratings as being on a nominal scale. We can then tabulate an assessment's results as shown in Figure 7. 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:

	Ass. 2				
Ass. 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 7: Notation for presenting proportions of ratings in each of the four rating categories by two assessors.

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

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

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

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

However, this value includes agreement that could have occurred by chance. For example, if the two assessors allocate ratings to process instances at random at the same rate as the marginal proportions (e.g., if assessor 2 in Figure 7 rated P_{+1} process instances as F, P_{+2} process instances as L, and so on) without regard to their adequacy in satisfying the purpose of the Generic Practice, 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 [10]:

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

Cohen [2] 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.

The value of Kappa depends strongly on the marginal distributions (see [11][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 instance. 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 [14]. Therefore, Kappa tends to be more conservative.

The variance of a sample Kappa has been derived by Fleiss et al. [12]. This would allow testing the one tailed alternative hypothesis that Kappa is greater than zero.

Given that we test 18 hypotheses for each of the 18 Generic Practices, the probability of incorrectly rejecting one of these null hypotheses (Type I error) is approximately 0.6. This means that there is reasonably high probability that at least one significant result would be found. We therefore use a Bonferroni adjusted alpha level for all hypothesis tests (see [19]).

3.4 Accounting for Seriousness of Disagreement

The above version of the Kappa coefficient assumes that all disagreements are equally serious. A weighted version of Kappa that allows different levels of seriousness to be attached to different levels of disagreement has been defined [3]. The weighted version of Kappa was used in a previous study on the reliability of process assessments [8].

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).

There are many potential weighting schemes that can be used. 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). This is a reasonable approach given that there is an implied ordering in the 4-point scale. This is the same approach that has been used in [8]. A suitable weighting scheme has been proposed by Fleiss and Cohen [13]:

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

where C is the number of categories, in this case 4.

3.5 Interpreting Interrater Agreement

After calculating the value of Kappa, the next question is “how do we interpret it?” We follow the guidelines developed and accepted within other disciplines. To this end, Landis and Koch [18] have presented a table that is useful and commonly applied for interpreting the obtained values of Kappa. This is shown in Figure 8.

It is still necessary to decide whether the obtained values of Kappa are good enough. Given the formative stage of research on the reliability of process assessments, it seems reasonable to consider Moderate agreement as a lower bound.

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 8: The interpretation of values of Kappa.

This is based on the argument that a lower bound should act as a good discriminator between assessments where the assessors had no difficulty making their ratings, and those where there was much uncertainty, misunderstanding and confusion about how to rate practices. It was deemed that Moderate agreement satisfied this requirement. However, future research that investigates the impact of disagreement on decision making would provide a stronger basis for identifying appropriate threshold values of agreement.

4 Results

4.1 Description of Assessments

The same two assessors conducted all of the assessments. Both assessors had relatively similar background profiles. They had on average 15.5 years of software employment experience each, have been conducting on average assessments for 5 years amounting to an average of 9 assessments each, with an average of 4 prior SPICE assessments each. Between them, they had conducted TickIT, SPICE, Bootstrap, and CMM-based assessments. A summary description of the projects assessed during the two assessments is given in Figure 9.

The variation in the reliability of the Generic Practices was not small. This is shown in Figure 10. Of particular concern are the Generic Practices that tend to have low interrater agreement levels. This is explored further below.

4.2 Estimates of Agreement

The overall results of interrater agreement are presented in Figure 11. The figure shows the Kappa estimate and its interpretation. Also shown are the number of process instances (n) that contributed to the Kappa value. This varies because for some process instances certain Generic Practices were not rated. This is evident as the value of n tends to decrease with higher capability levels. The reason is that the assessors start making their ratings from level 1, and stop rating a process instance when it is obvious that higher level ratings will produce “Not Adequate” values.

It was found that all Kappa values were significantly larger than zero at an overall probability of Type I error of 0.05. The “Allocate Resources” Generic Practices had the lowest Kappa value. This is due to the characteristics of the data set, however. Most of the ratings for that process were concentrated in one cell of the 4x4 table (as in

Assessment #	Project #	Project Duration*	# Staff	Customer	Functionality	# Processes Assessed
1	1	3 yrs	5	External	New	6
1	2	1 yr	3	External	New	4
1	3	3 yrs	6	Internal	Modification	3
1	4	4 yrs	3	Internal	New	3
2	1	1 yr	15	Internal	Modification	3
2	2	1 yr	3	External	New	3
2	3	2 yrs	3	Internal	New	3
2	4	2 yrs	5	Internal	Modification	2
2	5	1 yr	4	External	Modification	3
2	6	1 yr	5	Internal	Modification	3

* This is estimated project duration for incomplete projects.

Figure 9: Characteristics of the assessed projects.

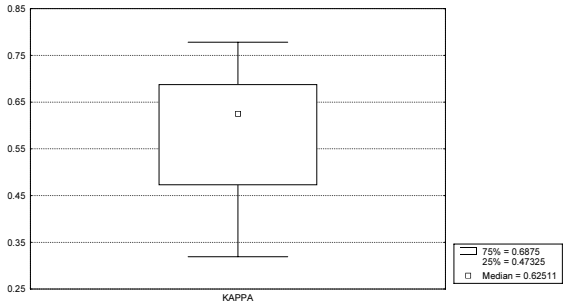

Figure 10: Variation in the value of Kappa.

Figure 7), therefore exhbiting little variation. This tends to attenuate the Kappa value even though percentage agreement may be high. The "Do Configuration Management" and "Perform Peer Reviews" Generic Practices also tended to have low Kappa values. This could be an indication that these two Generic Practices are not sufficiently well defined in the SPICE documents to enable highly reliable ratings. Except for "Do Configuration Management" and "Allocate Resources", all other Generic Practices were seen to have at least Moderate agreement in their ratings.

The results of applying our weighting scheme are shown in Figure 12. Recall that weighted Kappa assigns less weight to disagreements on adjacent categories. As can be seen, the values of Kappa do increase in general (compared to Figure 11). This is because most of the disagreements are on adjacent categories. All of the values of Kappa are at least Moderate. However, the "Perform Peer Reviews" agreement level is not statistically significant. Certainly the comparatively small sample contributes to that. Moreover, out of all disagreements, ratings for this Generic Practice disagreed approximately 23% on non-adjacent categories. This is a comparatively large number. Therefore, it seems that this particular Generic Practice is not well defined compared to other Generic Practices. Discussions with other assessors who have conducted assessments based on the SPICE framework confirmed that there is a general difficulty making ratings on this particular Generic Practice.

For the "Do Configuration Management" practice, there was disagreement on non-adjacent categories

Level	Generic Practice	n	Kappa	Interpretation
1	Perform the Process	33	0.68*	Substantial
2	Allocate Resources	32	0.32*	Fair
2	Assign Responsibilities	32	0.50*	Moderate
2	Document the Process	32	0.50*	Moderate
2	Provide Tools	32	0.47*	Moderate
2	Ensure Training	32	0.44*	Moderate
2	Plan the Process	31	0.69*	Substantial
2	Use Plans, Standards and Procedures	32	0.59*	Moderate
2	Do Configuration Management	32	0.40*	Fair
2	Verify Process Compliance	32	0.62*	Substantial
2	Audit Work Products	32	0.78*	Substantial
2	Track with Measurement	32	0.77*	Substantial
2	Take Corrective Action	32	0.75*	Substantial
3	Standardize the Process	23	0.68*	Substantial
3	Tailor the Standard Process	23	0.63*	Substantial
3	Use a Well-Defined Process	23	0.69*	Substantial
3	Perform Peer Reviews	23	0.41*	Moderate
3	Use Well-Defined Data	23	0.66*	Substantial

Figure 11: Kappa values for the 18 Generic Practices and their interpretations.

for only approximately 14% of all disagreements. Therefore, not surprisingly, the weighting scheme would increase the value of Kappa.

In summary, our findings indicate that ratings of the "Perform Peer Reviews" Generic Practice tend to have lower agreement. Ratings on the "Do Configuration Management" Generic Practice tend to also have low agreement, but disagreements tend to be on adjacent categories. While the results for the "Allocate Resources" indicate low agreement, but this is likely due to the characteristics of the data set used.

4.3 Limitations

There are two limitations to our current study that are discussed below. The discussions are intended to aid in interpreting our results, and also in identifying important issues for future reliability studies.

In interpreting our results, we have used a threshold of Moderate agreement. As alluded to earlier, this is not completely satisfactory as there is no strong empirical basis for this threshold. There are two obvious ways in which an improvement in the selection of the threshold can be made. The first is by comparison to results obtained using other assessment models and rating schemes. For example, a threshold can be set at the 50th pecentile of results obtained from studies with other models and rating schemes. However, despite the popularity of process assessments for more than a decade, the most extensive program of empirical research on the reliability of assessments has been conducted only recently in the context of the SPICE trials, and therefore there are no results from other models and rating schemes to compare with. The second approach is to define a threshold based on the cost of making erroneous decisions due to low reliability. However, a general cost function along these lines, to our knowledge, has not been developed.

Our conclusions came from a study using data from assessments conducted by two specific assessors. Whether these conclusions would hold for *any* two assessors remains a research question. The accumulation of results from multiple studies of interrater agreement would address this issue.

5 Conclusions

The objective of this paper was to evaluate the agreement between two independent assessors in their ratings of SPICE Generic Practices. Based on

Level	Generic Practice	n	Wtd. Kappa	Interpretation
1	Perform the Process	33	0.86*	Almost Perfect
2	Allocate Resources	32	0.53*	Moderate
2	Assign Responsibilities	32	0.84*	Almost Perfect
2	Document the Process	32	0.80*	Substantial
2	Provide Tools	32	0.67*	Substantial
2	Ensure Training	32	0.78*	Substantial
2	Plan the Process	31	0.89*	Almost Perfect
2	Use Plans, Standards and Procedures	32	0.82*	Almost Perfect
2	Do Configuration Management	32	0.74*	Substantial
2	Verify Process Compliance	32	0.77*	Substantial
2	Audit Work Products	32	0.93*	Almost Perfect
2	Track with Measurement	32	0.95*	Almost Perfect
2	Take Corrective Action	32	0.86*	Almost Perfect
3	Standardize the Process	23	0.83*	Almost Perfect
3	Tailor the Standard Process	23	0.74*	Substantial
3	Use a Well-Defined Process	23	0.83*	Almost Perfect
3	Perform Peer Reviews	23	0.56	Moderate
3	Use Well-Defined Data	23	0.78*	Substantial

Figure 12: Weighted Kappa values for the 18 Generic Practices and their interpretations.

this evaluation, we conclude that for most of the Generic Practices at Levels 1 to 3 there is at least Moderate agreement. We also identified a Generic Practice that exhibits lower agreement levels: *Perform Peer Reviews*. This practice should be the subject of future research to determine why it tends to have low levels of agreement in its ratings.

The current study focused on single assessor ratings. While this represents a sizeable proportion of assessments, especially in the context of small organisations, it would be of value to extend this kind of study to teams of assessors. In particular, comparing the interrater agreement between single assessor and team-based assessments.

Furthermore, the current study was limited to the first three levels of the capability scale used in SPICE. It is plausible that different conclusions would be drawn when studying the higher levels (levels 4 and 5) of capability.

Currently, the SPICE trials are on-going, and the interrater agreement of ratings using subsequent versions of the document set is being evaluated. It is planned that the reliability of team-based assessments and of higher capability levels will be evaluated during these empirical trials.

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

[4] A. Coletta: “Process Assessment Using SPICE: The Assessment Activities”. 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.

[5] 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.

[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, L. Briand, and R. Smith: “Assessor Agreement in Rating SPICE Processes”. In *Software Process Improvement and Practice Journal*, 2:291-306, 1996.

[8] 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, 1996.

[9] 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.

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

[11] A. Feinstein and D. Cicchetti: “High Agreement but Low Kappa: I. The Problems of Two Paradoxes”. In *Journal of Clinical Epidemiology*, 43(6):543-549, 1990.

[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] 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.

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

[15] ISO/IEC JTC1/SC7: “Software Process Assessment Part 4: Guide to Conducting Assessments”. 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] ISO/IEC JTC1/SC7: “DTR 15504-3: Information Technology - Software Process Assessment Part 3: Performing An Assessment”. Draft Technical Report, 1997.

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

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

[20] T. Rout and P. Simms: “Introduction to the SPICE Documents and Architecture”. 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.