Modelling the Reliability of SPICE Based Assessments

KHALED EL EMAM
BOB SMITH
PIERFRANCESCO FUSARO

International Software Engineering Research Network technical report ISERN-97-11

Modelling the Reliability of SPICE Based Assessments

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

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

Abstract

An area of major investigation in the SPICE trials is the reliability of assessments. In particular, their evaluation and improvement. Previous reliability studies in the trials have focused on evaluating reliability. In this paper we report on a study that focused on generating recommendations for improving the reliability of assessments through the construction of a model. The study attempted to identify factors that have an impact on reliability. Using data from three assessments, we constructed a model that explains some of the variation in the reliability of assessments. The two factors considered were the capability of processes and when ratings were made during an assessment (other factors such as assessor experience were held constant). Our model suggests that future assessment processes should have two contiguous phases in order to increase reliability. The first phase focuses only on data collection. During the second phase, ratings are made.

1 Introduction

The objective of the SPICE (Software Process Improvement and Capability dEtermination) Project is to deliver an ISO standard on Software Process Assessment. Unique to software engineering standardisation efforts, the SPICE Project includes a set of empirical trials [3][5]. 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 [3]).

The reliability of software process assessments has been studied in the past. Two types of reliability studies had been conducted: internal consistency [4][10] and interrater agreement [7][6]. Internal consistency is concerned with the extent to which components of an assessment instrument have been constructed to the same or consistent content specifications of what the instrument is supposed to measure. Initial internal consistency studies have been conducted within the SPICE trials [15][9]. Interrater agreement is concerned with the extent of agreement in the ratings given by independent assessors to the same organisational practices after being presented with the same evidence.

Until now, the focus of interrater agreement studies in the SPICE trials has been *evaluation*. This means that they aimed to evaluate whether the assessments are reliable or not. In general, the results for interrater agreement are encouraging but equivocal, with some processes not meeting minimal reliability thresholds [6].

Given such results, further studies are needed that will lead to improvements in interrater agreement. To this end, the objective of the research reported in this paper was to identify the factors that have an impact on the interrater agreement of SPICE-based assessments. This type of investigation has two benefits. First, it helps understand why some processes have high interrater agreement scores and others do not. Second, this understanding can be used to formulate recommendations for improving the reliability of assessments.

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 1: 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 2: Brief description of the capability levels.

The two factors that we investigated in this study are: (a) when ratings are made during an assessment, and (b) the capability of the process being assessed. Briefly, our results indicate that for low capability processes, ratings done later on in an assessment tend to be more reliable than ratings made early in the assessment. For high capability processes it does not matter when ratings are made. These results serve as a basis for making recommendations to improve SPICE-based assessment methods.

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 our study, and a description of the variables that we considered. Section 3 presents our research method, including data collection, measurement of the variables, and the analysis method. In section 4 we present the overall results. We conclude the paper in section 5 with a summary and directions for future work. The appendix includes the results of a survey that prioritised the factors that have an impact on the reliability of assessments.

2 Background

In this section we first 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. Since the completion of these assessments, version 2.0 of the SPICE documents has been released. However, the general conclusions we draw from our study should remain applicable to the version 2.0 documents as well. We then present the components of our model that explains variation in the interrater agreement of SPICE ratings.

2.1 Rating Scheme in SPICE v1.0

The SPICE architecture is two dimensional. 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 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 2.

Initially each base practice within a process is rated to determine the extent to which the process is actually performed. Once this 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 3. The four values are also designated as F, L, P, and N.

2.2 Factors Affecting Interrater Agreement

There are potentially a large number of factors that have an impact on the reliability of SPICE-based process assessments. Some of these may be related to the SPICE documents themselves, e.g., the clarity of the definition of the base and generic practices. Other factors may be related to the way the assessment was conducted, e.g., when the ratings were made during an assessment, or related to the people who are conducting the assessment, e.g., their skills and experience. In particular, such factors may influence the extent to which two independent assessors agree in their ratings of a process after being presented with the same evidence.

In constructing a model to explain variation in the reliability of assessments, we must first identify the factors that need to be considered. We then need to select the factors that would be allowed to vary and which ones to hold constant. Deciding on the subset of factors to vary is a compromise taking into account the available resources to perform the study.

The appendix of this paper describes a survey of experienced assessors that was conducted to prioritise the factors that have an impact on the reliability of assessments. The results of this survey are useful for interpreting the results of the current study. In particular, in the current study we hold many of the factors that were identified as important in the survey constant. Therefore, these factors would not influence variation in the reliability of the ratings made. Details of these factors are given in the appendix. The factors allowed to vary are described below.

The two factors that were included in our model (i.e., they varied in our study) are: TIME and CAPABILITY. As well as considering the impact of each of these factors directly on reliability, we consider interactions between them.

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.

2.2.1 TIME

This is the time at which the ratings were made. An assessment usually lasts for a few days. Reliability may differ between processes rated at the beginning of the assessment versus those rated at the end of the assessment. An increase in reliability over time indicates that the assessment team is gaining a better understanding of the organisation and of each other, hence a convergence in their ratings. If this is found to be the case then it would be recommended that, instead of ratings being made for each process right after information is collected about it, to collect information about all of the processes being assessed and then make the ratings at the end of the assessment. A decrease in reliability over time possibly indicates fatigue amongst the assessors, especially for long assessments. If this is found to be the case, then it would be recommended to set a limit on the number of days that an assessment can last.

An earlier study that evaluated the time effect [6] did not find a difference between the reliability of ratings done early in the assessment versus late in the assessment. However, this may have been due to the small sample size used in that study, justifying further investigation of the time factor.

2.2.2 CAPABILITY

We hypothesise that there is a relationship between process capability (i.e., capability of the process being assessed) and reliability. Making discriminations on the 4-point scale for processes with higher capability may be easier because they are more stable and there is likely to be more documented evidence to make fine judgements about their implementation. For lower capability processes, the instability and lack of documented evidence may cause the fine judgements on a 4-point scale more difficult to make and hence increase disagreement. Conversely, in [7] it was hypothesised that higher level generic practices are less reliable than lower level generic practices. This may be due to there being less general knowledge about the implementation of the high capability level generic practices.

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. 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, but can also be evaluated a posteriori. 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 2 groups would meet to reach a consensus or final assessment team rating (this is the set of ratings presented to management). General guidelines for conducting interrater agreement studies are given in Figure 4.

3.2 Data Collection

The data for this study was obtained from three assessments conducted within the European trials region during 1996. In total, 50 process instances were assessed.

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. 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 4.

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 ratings. Information was gathered for each process to be assessed through interviews and document reviews. Right after, the ratings for that process were made independently by the two assessors, and then the harmonised ratings are made. This is followed by a half day preparation of the final ratings

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 (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 to consensus building and harmonisation.

Figure 4: Guidelines for conducting interrater agreement studies.

and a meeting with the assessment sponsor. The assessment is then closed by a 2 hour presentation of the results of the assessment.

3.3 Measurement

Three variables had to be measured to test the model posited. The measurement of each of these is described in detail below.

3.3.1 Measurement of Interrater Agreement

To evaluate interrater agreement, 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:

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

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

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 teams employed completely different criteria for assigning their ratings to the same practices (i.e., if the row variable was independent from the column 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:

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

The above marginal proportions are maximum likelihood estimates of the population proportions under a multinomial sampling model. If each of the assessors makes ratings at random according to the marginal proportions, then the above is chance agreement (derived using the multiplication rule of probability and assuming independence between the two assessors).

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

	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.

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

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 κ=1. If observed agreement is greater than chance, then κ>0. If observed agreement is less than would be expected by chance, then κ<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.

3.3.2 Measurement of Capability

For each process instance assessed, a rating on the 4-point scale is given for each generic practice within the scope of the assessment. For example, if a process instance is rated up to Level 2, then ratings on 13 generic practices are made. In our analysis we have to convert these 13 ratings into a single number for that instance. This number would represent its capability.

We first assigned each of the possible 4 ratings a weighting. This weighting is based on the recommendation given in Part 4 of the SPICE document set [11], and reflects a consensus in the perceived weights that should be given to each ratings. This scheme assigns a value of 1 to an ‘F’ rating, 0.75 to ‘L’, 0.25 to ‘P’, and 0 to ‘N’. In calculating capability, we considered only the consensus/harmonised ratings for each process instance. We then summed up the weights assigned to the harmonised ratings for all rated generic practices for each process instance to get a capability rating.

One assumption made by this approach to capability calculation is that, when a process instance is not rated above a certain capability level then the likely rating for non-rated generic practices would have been Not Adequate. For example, if a process instance is only rated up to the Level 3 generic practices, then this reflects the opinion of the assessors that any ratings of generic practices above level 3 would produce ‘N’ ratings.

3.3.3 Measurement of Rating Time

For each process assessment, we dichotomised the projects by the time the rating was done. Two categories were used: Early and Late. *Early* was designated for ratings done earlier on in the assessment (approximately during the first half), and *Late* ones were done towards the end of the assessment (the second half of the assessment).

3.4 Data Analysis Method

The method that we used for analysing the collected data was multiple ordinary least squares regression with an interaction term. The general form of the regression model is:

$$\text{KAPPA} = b_0 + (b_1 \times \text{TIME}) + (b_2 \times \text{CAPABILITY}) + (b_3 \times \text{TIME} \times \text{CAPABILITY})$$

where:

KAPPA = the coefficient of interrater agreement
TIME = time of rating
CAPABILITY = the capability of the process

Since TIME can take only two values, it was treated as a dummy variable in the regression model and coded 0 for Early and 1 for Late.

This type of model allows us to investigate the main effects of both the independent variables, as well as their interaction. In particular, we assume that both TIME and CAPABILITY have a direct impact on KAPPA. Also, we assume that the effect of CAPABILITY on KAPPA depends on whether the ratings were done early or late in the assessment.

The analytical procedure that we followed is described in detail in [12]. This allows us to answer three questions: (1) “is there an interaction effect?”, (b) “if so, what is the strength of the effect?”, and (3) “if so, what is the nature of the effect?”. To answer the first question we consider if the regression coefficient of the interaction term is greater than zero. To answer the second question we compare the R^2 values of the model without the interaction term with the model with the interaction term (this

Assessment #	Project #	Project Duration*	# Staff	Customer	Functionality	# Processes Assessed
1	1	1 yr	4	External	New	5
1	2	1.5 yrs	4	Internal	New	5
1	3	1 yr	3	External	New	3
1	4	2.5 yrs	4	External	New	4
2	1	3 yrs	5	External	New	6
2	2	1 yr	3	External	New	4
2	3	3 yrs	6	Internal	Modification	3
2	4	4 yrs	3	Internal	New	3
3	1	1 yr	15	Internal	Modification	3
3	2	1 yr	3	External	New	3
3	3	2 yrs	3	Internal	New	3
3	4	2 yrs	5	Internal	Modification	2
3	5	1 yr	4	External	Modification	3
3	6	1 yr	5	Internal	Modification	3

* This is estimated project duration for incomplete projects.

Figure 6: Characteristics of the assessed projects.

gives us the increase in explained variation by adding the interaction term). To answer the third question, we calculate the two straight line equations for Early and Late assessments, and determine whether the slope is different from zero. The two equations are as follows:

$$KAPPA_{Early}=b_0 + (b_2 \text{ x } CAPABILITY)$$

$$KAPPA_{Late}=(b_0 + b_1) + ((b_2 + b_3) \text{ x } CAPABILITY)$$

The alpha level that we used for our statistical analysis was 0.1.

As noted in [12], the dummy variable should not be centred. We did not identify substantial multicollinearity effects, and therefore we did not centre the CAPABILITY variable either.

4 Results

4.1 Description of Assessments

Four assessors conducted all of the assessments: A, B, C, and D. Assessors A and B were the independent assessors for the first assessment, and assessors C and D were the independent assessors for the second and third assessments. A summary description of the projects assessed during the three assessments is given in Figure 6.

The variation in the reliability of assessed processes was quite substantial. This is shown in Figure 7 for the 4 process categories covered during the assessments. Hence, further justifying the construction of models to explain this variation.

To investigate for ability effect, we compared the interrater agreement scores obtained by the two assessment teams. A box and whisker plot showing the overall and interquartile ranges for the two teams is given in Figure 8. It is evident from this diagram that there is no substantial difference in interrater agreement between the two teams: the medians are quite similar and there is quite a large overlap in their variation. A more formal comparison, the t-test for independent samples, revealed no differences either at the 0.1 alpha level[1]. This is not surprising given that the two assessment teams had substantial and similar software engineering and assessment experience. A summary of the teams' experience is shown in Figure 9.

[1] We also tested for differences using the Mann-Whitney U test [14], which makes less assumptions than the t-test. Our conclusions would not change.

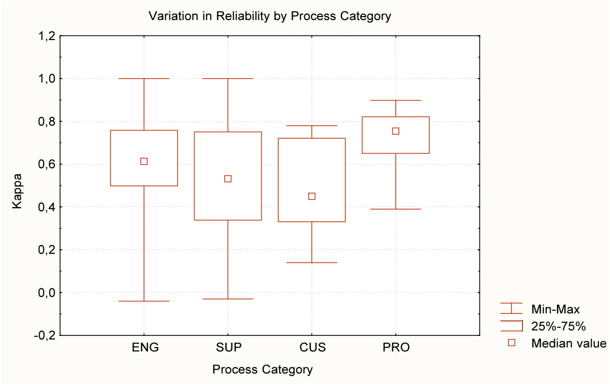

Figure 7: Variation in reliability scores.

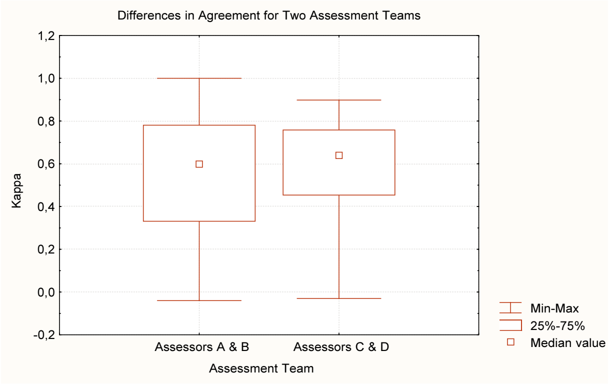

Figure 8: Effect of assessor experience.

Experience	Assessors A & B	Assessors C & D
Software Employment	11 yrs. (avg.)	15.5 yrs. (avg.)
Software Assessments	3 yrs. (avg.)	5 yrs. (avg.)
Number of Assessments	6.5 (avg.)	9 (avg.)
# of SPICE Assessments	3.5 (avg.)	4 (avg.)
Assessment Approaches Used	TickIT/SPICE/ CMM-based/ Bootstrap	ISO 9001 audits/ CMM-based/ Bootstrap

Figure 9: Summary of assessment teams' experience.

4.2 The Model

The resultant interaction model is given below[2]. The coefficients with an asterisk (*) were statistically significant at the 0.1 alpha level. The R^2 value was 0.31 and statistically significant.

$$\text{KAPPA} = 0.332^* + (0.504^* \times \text{TIME}) + (0.020^* \times \text{CAPABILITY}) - (0.038^* \times \text{TIME} \times \text{CAPABILITY})$$

We now revisit the three questions we raised earlier about this interaction model. First, the regression coefficient for the interaction term is statistically significant. Therefore, there is an interaction effect. Second, the difference in R^2 values between the model without the interaction term and the interaction term was 0.31 - 0.17, giving a 14% increase in explained variation with the addition of the interaction term. This value is comparable to that obtained in a previous software engineering study [8].

The two equations for Early and Late ratings are:

$$\text{KAPPA}_{\text{Early}} = 0.332 + (0.020 \times \text{CAPABILITY})$$

$$\text{KAPPA}_{\text{Late}} = 0.836 - (0.018 \times \text{CAPABILITY})$$

For the Early equation, the slope coefficient was statistically significant. For the Late equation, the slope was not statistically significant. This means that getting a slope as different from zero as this one could happen by chance.

4.3 Discussion

The results of this study identify a number of explanations for the variability in the interrater agreement of SPICE processes. First, there is a difference in reliability between ratings done early in the assessment versus ratings done later on in the assessment. This is evident by the fact that Kappa values of Late ratings tend to be higher than for Early ratings for most of the capability values in our sample: the intercept for the Late equation is more than twice the intercept for the Early equation. Furthermore, the reliability of Late ratings does not change with changes in the capability values.

[2] In this analysis we had to remove a number of data points because either the lack of variation in ratings gave extreme Kappa values or because the observations were considered to be outliers. In total 4 data points had to be removed for the first reason, and 3 for the second reason, leaving an n of 43.

On the other hand, Early ratings tend to be affected substantially by the capability of processes being assessed. For low capability processes, Early ratings tend to have low reliability. This is interpreted to be due to rating on a 4-point scale that requires greater discrimination than the assessors are able to make (and hence they disagree more), especially when they have not yet spent time understanding the other processes in the organisation. For low capability processes there also tend to be less work products that assessors can use to inform their rating judgements. As the capability of processes increases, the discrimination ability of the assessors increases, even though they have not looked at all of the other processes within the scope of the assessment.

These results have a number of implications for improving the reliability of SPICE-based assessments. It is generally more advisable to rate processes later on in the assessment, as this ensures more reliable ratings irrespective of the processes' capabilities. This suggests a two phase assessment process, where assessors first collect information and then make ratings after all of the necessary information has been collected. For higher capability processes, it generally does not matter whether ratings are made early or late in the assessment. However, given that in general few processes have high capability (e.g., see [16]), the two phased assessment process is suggested for contemporary assessments.

Our results also indicate that the assessors' ability to discriminate, depends on both, the capability of the processes and when the ratings are actually made. Therefore, general statements about the appropriateness of the 4-point scale remain premature.

The model we constructed accounted for only 31% of the variation in the reliability of ratings. This indicates that there are other factors that need to be considered in order to account for the remaining unexplained variance. From the results of the survey in the appendix, the most important factors are worth revisiting in this context. In particular, assessor experience and knowledge of the SPICE/WG10 documents. In our study these were above the required minimum stipulated in the SPICE documents, and the four assessors also tended to have similar experiences. However, perhaps there are other assessor characteristics that were not considered and that did vary in our study, or that characteristics of finer granularity need to be evaluated (e.g., knowledge of SPICE documentation on specific process categories). This issue deserves further empirical investigation. A second issue deserving of further investigation is the clarity and semantics of the process definitions in the SPICE documents. When considered at the overall SPICE document level, this was constant in our study. When considered at the level of a process category, this factor does vary amongst different process categories, and therefore the explicit consideration of process categories in a model may improve the amount of explained variation.

4.4 Threats to Validity

One alternative interpretation of the results that we have obtained is that rating some process categories is more reliable than others, and this may have influenced our results. This would be the case if certain process categories tended to be rated Early or Late, which may be the reason why the reliability is different between the two groups. For example, if we assume that processes of the ENG category can be rated more reliably than other process categories, and if we find that there was a greater probability of finding an ENG process in the Late group (as opposed to the Early group), then we would expect this to be a potential cause of finding that the Late group has more reliable ratings than the Early group. To determine if this was the case, we evaluated whether there was a relationship between TIME and the four different process categories that were assessed in our study.

We used Fisher's exact test for a 4x2 table to conduct the test. The network algorithm described by [13] was used to calculate the probabilities. The test computes the probability of finding 4x2 tables more extreme than the current table, holding the marginals fixed.

The p value was found to be larger than 0.5, and hence highly non-significant. We therefore found no association between TIME and process category. This is justification for dismissing process category as the reason for the difference in reliability between Early and Late assessments.

For both pairs of assessors, it was the first time that they have worked together on an assessment. This may have introduced some variation in reliability between the first and second assessments by assessors C and D.

Finally, it was taken for granted that the measure of capability we used was valid. This measure was based on formulations in the SPICE documents. However, to our knowledge, there is no published empirical evidence that this measure really measures process capability.

4.5 Post-hoc Analysis

In our study all assessors were external. There are two possible variations to this makeup of an assessment team. One is to have the team consist of only internal assessors, and the other is to have a mixture of internal and external assessors. A previous study [6] had indicated that there may be systematic bias by either an internal or external assessor. For example, an internal assessor may consistently rate his/her organisation favourably or an external assessor may be consistently too critical of an organisation's practices. Systematic bias reduces reliability.

We wanted to determine whether such systematic biases also existed if the assessment team consists solely of external assessors, as is the case in our study. If such is the case, then there is evidence that composition of the team is not a factor, and that systematic bias is caused by something else.

To test for this possibility, we used the sign test [14]. For each assessed process, we assign a '+' if the first assessor's ratings is greater than the second, and a '-' if the it is less. We ignore ties. Under the null hypothesis, we would expect that the number of +'s to be the same as the number of -'s.

We found that six of the fifty processes assessed demonstrate systematic bias. This result indicates that also when the assessment team consists of only external assessors, systematic bias could occur. This indicates that systematic bias is not necessarily only a function of the team composition.

5 Conclusions

In this paper we constructed a model to explain the variation in the reliability of assessing SPICE-based processes. The explanatory factors that were examined were the capability of processes being assessed and when ratings are made during an assessment. Data from three assessments were used to construct the model.

The results indicate that for low capability levels, there is a difference in reliability between rating processes early in assessment versus late in an assessment. For higher capability processes, it does not make a difference whether ratings are done early or late in an assessment.

These results suggest that a two phased assessment process is advisable for increasing the reliability of assessments. The first phase consists of only data collection, and second phase consists of making the actual ratings.

Future empirical studies in the SPICE trials will attempt to investigate other factors that have an impact on the reliability of assessments. This would lead to further recommendations for improving the reliability of assessments.

Acknowledgments

The authors wish to thank Jerome Pesant for collecting the survey data.

References

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

[2] K. Dymond: "Essence and Accidents in SEI-style Assessments or 'Maybe This Time the Voice of the Engineer will be Heard'". In *Software Process Newsletter*, IEEE TCSE , pages 1-7, No. 9, Spring 1997.

[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 Organisational 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):123-148, 1996.

[6] K. El Emam, L. Briand, and R. Smith: "Assessor Agreement in Rating SPICE Processes". To appear in *Software Process Improvement and Practice Journal*, 1997.

[7] 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.

[8] K. El Emam and N. H. Madhavji: "User Participation in the Requirements Engineering Process: An Empirical Study". In *Requirements Engineering Journal*, 1:4-26, 1996.

[9] P. Fusaro, K. El Emam, and B. Smith: "The Internal Consistencies of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension". Technical Report, International Software Engineering Research Network, ISERN-97-01, 1997.

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

[11] ISO/IEC JTC1/SC7: "Software Process Assessment Part 4: Guide to Conducting Assessments". Working Draft 1.0, 1995.

[12] J. Jaccard, R. Turrisi, and C. Wan: *Interaction Effects in Multiple Regression*. Sage Publications, 1990.

[13] C. Mehta and N. Patel: “A Network Algorithm for Performing Fisher’s Exact Test in rxc Contingency Tables”. In *Journal of the American Statistical Association*, 78:427-434, 1983.

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

[15] R. Smith and K. El Emam: “Transitioning to Phase 2 of the SPICE Trials”. In the *Proceedings of the SPICE’96 Conference*, pages 45-55, 1996.

[16] 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.

Appendix: Prioritising Factors Affecting Reliability

We conducted a survey to prioritise the factors that have an impact on the reliability of process assessments. The results from this survey are helpful in interpreting the results of the study that we report in the body of the paper, and also for highlighting future avenues for research on the reliability of assessments.

The data collection was conducted during a meeting of the SPICE project that took place in Mexico in October 1996. These meetings are of sizeable number of experienced assessors with substantial experience in various assessment methods and models, such as the CMM, CBA IPI, TRILLIUM, BOOTSTRAP, and other proprietary models and methods. During the meeting, the authors generated a list of factors that may potentially have an impact on the reliability of assessments. We relied largely on our experiences and the prior comments of other assessors. This is justifiable given that no comprehensive study of the factors influencing reliability has been conducted thus far, and therefore our list could serve as a starting point for subsequent studies.

This list was reviewed by two other experienced assessors to ensure completeness of coverage. The list is given in Figure 10. The refined list was turned into a rating form. The rating form was piloted with 4 assessors to ensure that it was understandable and to identify ambiguities. Based on this feedback, a new form was developed and was distributed to all attendees at the closing session of the meeting. In total, approximately 50 individuals attended the project meeting, and we expect a slightly smaller number attended the closing session. We received 26 valid responses back. These are the responses that we used in this analysis.

The form consisted of an unordered list of factors that are believed to have an impact on the reliability of assessments. The respondent was requested to rate each factor on a five point scale, where 1 means that the factor has "very high influence" on the reliability of assessments, and 5 means that it has "very low influence". Our objective was to prioritise these factors. So we dichotomised the responses on the 5-point scale into HIGH INFLUENCE (scores 1 and 2) and LOW INFLUENCE (scores 3 to 5). For each factor, we then calculated the percentage of respondents who rated a factor as HIGH INFLUENCE. This percentage is used for ranking.

It should be noted that these results are not indicating issues that are suboptimal with SPICE, but the ones that are believed to be important from the perspective of reliability of assessments in general.

Results

The results are presented in Figure 10. Below we discuss these results and consider their impact on the study reported in the body of the paper. We use the letters A to W in the discussions to indicate items in Figure 10.

Assessor Competence

Given that assessment are a subjective measurement procedure, it is expected that assessor competence will have an impact on the reliability of assessments. We consider mainly the competence of the lead assessor since s/he is the key person on the assessment team. The types of competences covered here include knowledge of the SPICE documents (B), experience and competence in conducting assessments (A) and audits (P). The distinction between assessments and audits is important because the manner in which an assessment is conducted does differ between the two, an audit being more adversary. For the remainder of the assessment team, we consider their knowledge of the SPICE documents (E) since the team members will be collecting, organising, and interpreting information during an assessment, they must know SPICE well to collect the right information, organise it efficiently, and interpret it properly.

All of the assessors that took part in our study were experienced and met the minimal qualification

Id	Factor	Percentage Who Think it is Important
A	Lead assessor's experience/competence in conducting assessments	(24/26) = 92%
B	Lead assessor's knowledge of SPICE or the WG10 documents	(22/25) = 88%
C	Clarity of the semantics of the process definition in the SPICE or WG10 documentation	(22/26) = 84.6%
D	The extent to which the assessment process is defined and documented	(20/26) = 77%
E	Team members' knowledge of SPICE or the WG10 documents	(16/25) = 64%
F	Amount of collected data (objective evidence and/or interviews)	(16/26) = 61.5%
G	Assessees' commitment	(13/25) = 52%
H	Assessment team stability	(13/25) = 52%
I	Rating just after collecting the evidence, and validation at the end of the assessment	(12/25) = 48%
J	Assessment team composition (unidisciplinary vs. multidisciplinary competences)	(11/25) = 44%
K	Sponsor commitment	(11/25) = 44%
L	Team building curve	(10/25) = 40%
M	Competence of the interviewed assessees	(10/25) = 40%
N	Number of assessed projects in the organisational unit	(9/25) = 36%
O	Assessment duration	(9/25) = 36%
P	Lead assessor's experience/competence in conducting audits	(8/25) = 32%
Q	Assessment team size (number of assessors including lead assessor)	(8/25) = 32%
R	Rating only at the end of the assessment	(8/25) = 32%
S	Language used during the assessment	(8/25) = 32%
T	Time allocation between artifact reviews and interviews	(8/25) = 32%
U	Management of the assessment logistics (e.g., availability of facilities)	(6/25) = 24%
V	The capability of the organisational unit's processes	(5/25) = 20%
W	Whether the assessors are external or internal	(3/26) = 11.5%

Figure 10: Factors affecting the reliability of assessments and their prioritisation.

requirements for conducting a SPICE-based assessment. There was no variation in this factor, and therefore its impact on the variation in reliability is minimised. 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 required to be performed by qualified assessors; not having qualified assessors is not relevant for evaluating SPICE conformant assessments[3].

External vs. Internal Assessors

Previous research has identified potential systematic biases of internal or external assessors [6] (i.e., one assessor would systematically rate higher or lower than the other). For example, an internal assessor may favour the organisation in his/her ratings or may

[3] Although it may be interesting for evaluating the conformance criteria, this was not our objective here.

have other information not available to the external assessor which may influence the ratings. Similarly, an external assessor may not know the organisation's business well and therefore may systematically underrate the implementation of its practices. This issue is covered in item (W).

All assessors who took part in our study were external. This removes the above source of bias.

Team Size

Practice and recommendation on team size have tended to be confusing. In some assessment methods it is stipulated that teams range in size from 5 to 9 [2]. In the first version of the SPICE documents the recommendation has been team sizes of at least two assessors. In the first phase of the SPICE trials approximately 9% of the assessments had one-person assessment teams [6]. The second version of the SPICE documents allows team sizes of one, especially for small assessments. From a practical point of view it has been suggested that a single assessor would find it difficult to collect and record information at the same time, and therefore more than one person is recommended. Item (Q) covers this issue from the perspective of its impact on reliability.

In the current study independent single assessors performed the ratings. Single assessor assessments have the advantage of lower cost, which makes the assessment more attractive for small organisations. Therefore studying them is of practical utility.

Backgrounds of Assessors

It has been noted by some assessors that multidisciplinary assessment teams (i.e., not consisting of only software engineering staff, but also those with backgrounds in, for example, human resources management and marketing) are better able to collect the right evidence (i.e., ask the right questions and request the appropriate documents) and better able to interpret it for certain processes. This would likely increase the reliability of assessments. This issue is covered in item (J).

This factor did not vary in our study.

Number of Assessed Projects

During an assessment, a sample of projects is selected for assessment. It is usually not feasible to assess all of the projects within the scope of the organisation. It is assumed, through this selection process, that the selected projects are representative of the whole organisation. Clearly, the more projects that are assessed, the more representative and hence repeatable the ratings that are made. This is covered in item (N). Of course, this item applies only when one is giving ratings to whole organisations, and has less influence when the unit of analysis is a process instance.

Assessment Duration

Long assessment may lead to fatigue of the assessors and assessees, may reduce their motivation, and hence reduce the reliability of ratings. Short assessments may not collect sufficient information to make reliable ratings. This is covered in item (O). This factor did not vary in our study.

Team Building Curve

In team-based assessments, it is expected that the assessor judgements would converge as the assessment progresses. This would be due to a better appreciation of the other team members' experiences, backgrounds, and due to the consensus building activities that usually take place during an assessment. This is covered in L.

This point is discussed further in the section on threats to validity as it pertains to our study.

Clarity of Documents

Ambiguities and inconsistencies in the definition of practices or in the scales used to make ratings would potentially lead to different interpretations of what practices mean and how to rate them. This would in turn reduce reliability. This issue is covered in C. Since the same documents were used as the basis of all assessments, this factor did not vary in our study.

Definition of the Assessment Process

Having a clearly defined assessment process potentially ensures that the the process is repeatable, which in turn has an impact on the repeatability of ratings. This is covered in D.

In our study, all assessments followed the same process and is documented in the training material available to assessors.

Amount of Data Collected

The more time spent on data collection (F), the more data will be collected. The more data that is collected, the more likely that the assessment team will have a more objective basis to make their ratings. Furthermore, the extent to which time is allocated to different methods for data collection may have an impact on the amount of data collected (T).

In our study, similar amounts of data were collected for the different process instances assessed.

Capability of Organisation and its Processes

It is hypothesised that higher capability processes are easier to rate because of the existence of more objective evidence and process stability to make consistent judgements. This is covered in U. This factor varied in our study.

Assessment Method

A feature of the assessment method is *when* the ratings are actually made. One approach is to collect data about a process and then make the ratings right afterwards (I). Another approach is to collect data on all of the processes within the scope of the assessment, and then rate them all afterwards (R). The latter allows the assessors to build an overall picture of the implementation of software engineering practices in the organisation, and also to get a better understanding of the organisation's business and objectives (especially for external assessors) before making ratings. This could potentially increase the reliability of assessments.

This is one of the factors considered in our study.

Sponsor and Assessee Commitment

A lack of commitment by members of the assessed organisation can lead to insufficient or inappropriate resources being made available for the assessment. This may compromise the assessment team's ability to make repeatable ratings. This issue is covered in items (K and G).

This factor did not vary in our study.

Assessment Team Stability

If the assessment team changes during an assessment, the disruption can break the consensus building cycle. Furthermore, knowledge about the organisation that has been gained by an assessor that leaves would have to be regained by a new assessor. This is covered in item H.

This factor did not vary in our study.

Logistics Management

Inappropriate management of the logistics may distract the assessors and waste time. This could potentially lead to insufficient evidence being collected and hence to lower reliability. This issue is covered in (V).

Assessee Competence

Assessees provide the necessary information during an assessment. If the assessees are not competent then they may provide inconsistent information to the assessors, which may consequently lead to inconsistent interpretations of the process' capability. This issue is covered in item (M).

Assessment Language

Assessments are now being conducted all around the world. In fact, in the first phase of the SPICE trials certain documents were translated to a language other than English. One of the aims of the SPICE framework is that it should be culturally independent. The issue of the impact of language on the reliability of assessments is covered in (S).

The language used in the assessments of our study was English and that did not vary.

Discussion

The results clearly indicate that assessment team competence and the clarity of the documents are the two most important factors that have an impact on the reliability of assessments. This has a number of implications for research and practice.

Equally interesting are the factors that were rated to be of least priority. This does not mean that they are not important, only that they are less important than the other factors. These factors were whether the assessors were internal vs.external, the capability of assessed processes, and the assessment logistics. However, the capability of assessed processes was found in the current study to have a nontrivial impact on reliability, indicating that these factors should also be studied.