Fraunhofer Einrichtung
Experimentelles
Software Engineering

Benchmarking Kappa for Software Process Assessment Reliability Studies

Authors
Khaled El Emam

IESE-Report No. 016.98/E
Version 1
May 1998

A publication by Fraunhofer IESE

Fraunhofer IESE is an institute of the Fraunhofer Gesellschaft.
IESE transfers innovative software development techniques, methods and tools into industrial practice, assists companies in building software competencies customized to their needs, and helps them to establish a competetive market position.

Fraunhofer IESE is directed by
Prof. Dr. Dieter Rombach
Sauerwiesen 6
D-67661 Kaiserslautern

Benchmarking Kappa for Software Process Assessment Reliability Studies

Khaled El EMAM

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

International Software Engineering Research Network Technical Report ISERN-98-02

Benchmarking Kappa for Software Process Assessment Reliability Studies

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

Abstract

Software process assessments are by now a prevalent tool for process improvement and contract risk assessment in the software industry. Given that scores are assigned to processes during an assessment, a process assessment can be considered a subjective measurement procedure. As with any subjective measurement procedure, the reliability of process assessments has important implications on the utility of assessment scores, and therefore the reliability of assessments can be taken as a criterion for evaluating an assessment's quality. The particular type of reliability of interest in this paper is interrater agreement. Thus far, empirical evaluations of the interrater agreement of assessments have used Cohen's Kappa coefficient. Once a Kappa value has been derived, the next question is "how good is it?" Benchmarks for interpreting the obtained values of Kappa are available from the social sciences and medical literature. However, the applicability of these benchmarks to the software process assessment context is not obvious. In this paper we develop a benchmark for interpreting Kappa values using data from ratings of 70 process instances collected from assessments of 19 different projects in 7 different organizations in Europe during the SPICE Trials (this is an international effort to empirically evaluate the emerging ISO/IEC 15504 International Standard for Software Process Assessment). This benchmark can be used to decide how good an assessment's reliability is.

1. Introduction

The reliability of software process assessments has received considerable empirical evaluation in the last three years. The most extensive program of research on this area has been conducted in the context of the SPICE Trials (see [24]). The SPICE Trials are being conducted to empirically evaluate the emerging ISO/IEC 15504 International Standard for Software Process Assessment in order to inform design decisions and also to provide guidance in applying the emerging standard. The strong focus on reliability is driven by the recognition that the utility of process assessments, in general, is enhanced the more reliable they are, and that quite erroneous decisions can be made by applying assessment scores obtained from unreliable assessments (see [6][7]).

There are different types of reliability that can be evaluated. For example, one type is the internal consistency of instruments (see [6][7][19]). 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 [9].

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 ISO/IEC 15504 rely on the judgement of experienced assessors in assigning ratings to software engineering practices. This means that there is an element of subjectivity in their ratings. Ideally, if different assessors satisfy the requirements of the ISO/IEC 15504 framework and are presented with the same evidence, they will produce exactly the same ratings (i.e., there will be perfect agreement amongst independent assessors). In practice, however, the subjectivity in ratings will make it most unlikely that there is perfect agreement. The extent to which interrater agreement is imperfect is an empirical question.

Our focus in this paper is on interrater agreement. High interrater agreement is desirable to give credibility to assessment results, for example, in the context of using assessment scores in contract award decisions. If agreement is low, then this would indicate that the scores are too dependent on the individuals who have conducted the assessments.

The statistic that has been employed almost exclusively in the evaluation of interrater agreement has been Cohen's Kappa [5]. For example, a series of interrater agreement studies in the SPICE Trials have used Kappa [13][26][8][10][11][18].

Outside software engineering, the Kappa statistic is quite popular with researchers for evaluating intra and inter observer agreement. For instance, Kappa has been used to evaluate the agreement in identifying mental disorders, such as depression, neurosis, and schizophrenia [15]. Umesh et al. [27] note that up to April 1988 Kappa had been cited more than 1100 times in social science research. This number is undoubtedly much larger by now. Furthermore, in medical methodology texts Kappa has been presented as a measure of agreement in diagnosis reliability studies [2][3][20].

As with any other statistic, after the computation of the value of Kappa, it is necessary to be able to decide whether the value obtained is good enough. This is of particular import for software process assessments since it has been shown in previous studies that Kappa values exhibit substantial variation even for assessments conducted using the same assessors and using the same assessment method [11]. With appropriate interpretation, the Kappa value can be used as an objective criterion for evaluating the quality of an assessment. Of course, given that the expression for the standard error of Kappa under certain assumptions has been derived [17], it is possible to perform hypothesis testing. However, it has even been suggested that hypothesis testing of Kappa is not needed because in practice agreement is usually better than chance anyway [3]. Therefore, if one accepts the latter argument, then the most important issue is to interpret the actual value of Kappa rather than hypothesis testing.

In the social and medical sciences a number of authors have proposed benchmarks for interpreting the value of Kappa (i.e., deciding how good agreement is). However, no benchmark has been developed for interpreting Kappa for software process assessment reliability studies. Using benchmarks from other disciplines directly may not be appropriate as these are based on experiences with completely different subjective measurement procedures. For example, if a medical diagnosis procedure is considered to have high reliability only if its Kappa value is greater than 0.8, then this may be too stringent in our context if very few software process assessments can actually achieve Kappa values that high in practice. The converse may also be true.

Given this state of affairs, it becomes important to develop a software process assessment benchmark for interpreting Kappa values. Such a benchmark can be used in a number of ways, for example, as a basis for evaluating the quality of an individual assessment, and as a basis for interpreting the results of empirical studies of interrater agreement.

The objective of this paper is to report on a study to construct a software process assessment specific benchmark for Kappa. We use data obtained from process assessments of 70 process instances during the SPICE Trials. The benchmark has four ranges or levels based on the quartiles of the Kappa distribution.

In the next section we present the background to our study, including a brief discussion of the general scheme for rating processes used in ISO/IEC 15504, an overview of existing benchmarks, and the different approaches to constructing a benchmark. This is followed in Section 3 by a description of the data sources. Section 4 presents our benchmark and a discussion of how to use it and its limitations. We conclude in section 5 with a summary and directions for future work.

2. Background

2.1 ISO/IEC 15504 Rating Scheme

Our study used data collected from assessments using Version 1.0 of the SPICE document set. The complete version 1.0 document set is available in [12]. Below we summarize the main characteristics of the SPICE architecture.

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*.

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

[1] Elements of the SPICE architecture have recently been revised and restructured (for example, see [12]). The basic two dimensional architecture remains however.

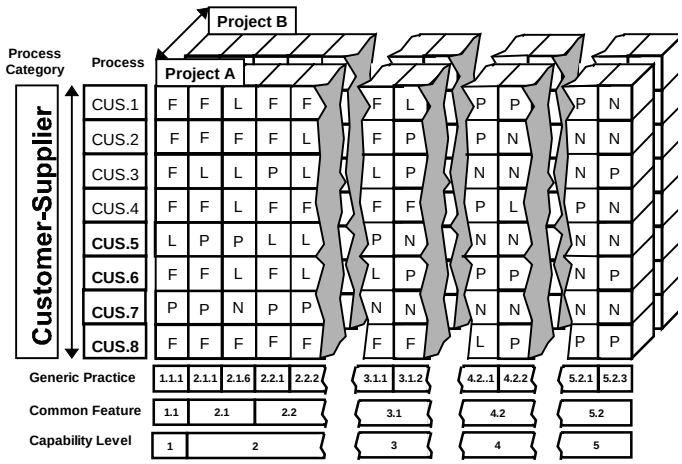

Figure 1: General architecture and rating scheme (source [28]).

This two dimensional architecture is illustrated in Figure 1 for the "Customer-Supplier" Process Category. Ratings are made at the intersection of the process and capability dimension. The unit that is rated is called a *process instance* [25]. A process instance is a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner. Process instances often correspond to software projects. A particular project will instantiate, for example, a set of management, support, and engineering processes. In Figure 1, there are two projects, A and B, that are being rated.

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

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 2: Description of the rating scheme for generic practices.

2.2 An Overview of Kappa

Data from an interrater agreement study of an ISO/IEC 15504 assessment can be represented in a table such as Table 1. Here we have two teams that have independently made a number of ratings on the 4-point scale described above (the design of such studies is described below). The table would include the proportion of ratings that fall in each one of the cells.

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), 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 [5] has defined coefficient Kappa (κ) as an index of agreement[2]. 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.

[2] 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.

		Team 1				
		F	**L**	**P**	**N**	
Team 2	**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+}
		P_{+1}	P_{+2}	P_{+3}	P_{+4}	

Table 1: 4x4 table for representing *proportions* from an ISO/IEC 15504 interrater agreement study.

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.

2.3 Existing Benchmarks

Hartmann [21] gives a basic benchmark for Kappa values: they should exceed 0.6. Landis and Koch [22] provided a more detailed benchmark for interpreting the values of Kappa. This is summarized in Table 2. Landis and Koch concede that their benchmark is arbitrary, but they nevertheless contend that it can serve as a useful guideline. The same point is echoed by Everitt [14]. Altman [2] presents a slightly modified version of the Landis and Koch benchmark. This is summarized in Table 3. Although, he does go on to say that any value of Kappa much below 0.5 would indicate poor agreement. Fleiss [16] has presented a slightly different benchmark, also based on Landis and Koch. This is summarized in Table 4. This benchmark has appeared in medical methodology texts, such as [3][20]. Presumably the acceptance of the Landis and Koch, Altman, and Fleiss benchmarks is a consequence of accumulated experience in using Kappa in medical studies, whereby the benchmarks were found to be useful.

Previous studies of interrater agreement of software process assessments have also used the Landis and Koch benchmark to interpret results [8][26][18]. This seemed reasonable at the time given the dearth of software engineering experiences with evaluating interrater agreement. However, as noted in [18], this is not completely satisfactory as this benchmark is not based on accumulated experiences in software engineering. For example, the Landis and Koch benchmark may be too optimistic compared to what can be realistically achieved using process assessments as a subjective measurement procedure.

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

Table 2: The Landis and Koch Kappa benchmark.

Kappa Statistic	Strength of Agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very Good

Table 3: The Altman Kappa benchmark.

Kappa Statistic	Strength of Agreement
<0.40	Poor
0.40-0.75	Intermediate to Good
>0.75	Excellent

Table 4: The Fleiss Kappa benchmark.

2.4 Constructing a Benchmark

In the measurement literature, the interpretation of measured values can be *norm-referenced* or *criterion-referenced* [1]. With a norm-referenced interpretation one would derive an interpretable score from a raw score by comparing them with values obtained from a reference sample. This would give an indication of how good a raw score value is with reference to what has been achieved in previous assessments. With criterion-referenced interpretation, one would compare the obtained value of Kappa with a threshold to determine whether it is above or below (i.e., better or worst). This threshold would be a critical Kappa value above which reliability is “good enough”, for example, for making a particular decision.

Since there is no clear general basis (e.g., an economic basis) for defining a Kappa threshold (or series of thresholds), the criterion-referenced approach would be difficult to operationalize. The norm-referenced approach can be operationalized using data from past studies. Since a considerable amount of data is available from the SPICE Trials that can be applied for constructing a benchmark, we follow a norm-referenced approach. Care should be taken in interpreting a norm-referenced benchmark, however, since the representativeness of the data used has a direct impact on how the results should be interpreted. This issue of interpretation will be discussed later in the paper.

One common technique for constructing norm-referenced scores is to use percentiles [23]. This has the advantage of deriving scores that are easily understandable, and it does not make distributional assumptions. Since we are only interested in defining ranges of Kappa to construct a benchmark, we only derive the 25%, 50%, and 75% percentiles. The same technique has been applied to construct a software inspection efficiency benchmark [4].

3. Data Sources

We use data from four interrater agreement studies that have been conducted thus far[3] [13][8][11][10]. All four studies followed the same guidelines for data collection. These are summarized below.

For conducting interrater agreement studies, the assessment team is divided into two or more groups. Ideally all groups should be equally competent in making attribute adequacy ratings. In practice, assessors in each group need only meet minimal competence requirements since this is more congruent with the manner in which the 15504 documents would be applied. 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) [4], and then they would perform their ratings independently. Subsequent to the independent ratings, the groups would meet to reach a consensus or final assessment team rating. General guidelines for conducting interrater agreement studies are given in Table 5.

[3] These are studies with different data sets *and* where sufficient data was collected to facilitate the computation of Kappa per process instance.

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

Instructions for Conducting Interrater Agreement Studies
• For each process, divide the assessment team into k>=2 groups with at least one person per group. • The k groups should be selected so that they both meet the minimal assessor competence requirements with respect to training, background, and experience. • The k groups should use the same evidence (e.g., attend the same interviews, inspect the same documents, etc.), assessment method, and tools. • Each 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 the k groups should inspect the new evidence before making ratings. • The k groups independently rate the same process instances. • After the independent ratings, the k groups then meet to reach consensus and harmonize their ratings for the final ratings profile. • There should be no discussion amongst the k groups about rating judgment prior to the independent ratings[5].

Table 5: Guidelines for conducting interrater agreement studies.

The above assessments were all conducted by professional assessors in the context of the SPICE Trials. In total, our data set represents assessments of process instances from a total of 19 different projects conducted in 7 different organizations in Europe. Three of the assessments had two independent assessors performing the ratings (k=2) with one assessor per group, and in one [13] there were three independent assessors (k=3) with one assessor per group.

In total, 75 process instances were assessed in these four studies. Four of these 75 produced data that were badly distributed and therefore were unusable (i.e., all ratings were concentrated in one cell, and hence the Kappa statistic cannot be computed). One observation was an outlier which was found to be due to an extreme case of bias by one of the independent assessors (see [8]). This left 70 process instances for which we had usable Kappa values.

4. Results

A box and whisker plot of the remaining 70 values of Kappa is given in Figure 3. This reflects the variation in Kappa obtained from actual process assessments. Such variation is expected due to, for example, different processes being assessed (some processes are more difficult to rate than others, and this affects the reliability), different process capabilities (higher capability processes tend to be more reliable [11]), as well as a myriad of other factors summarized in the appendix of [11]. When constructing a benchmark, it is desirable to capture this variation since it reflects actual assessments.

The benchmark using the upper quartile (0.78), median (0.62), and lower quartile (0.44) is given in Table 6. The table also gives linguistic descriptions for each one of the ranges. The interpretation of this benchmark is straight forward. If, for example, an interrater

[5] 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 different groups would no longer be independent. This can be addressed if consolidation is performed independently by the different groups. Then, before the presentation of draft findings to the organization, independent ratings are given followed by consensus building and harmonization of ratings by the different groups.

agreement study produces a value of Kappa of 0.85, then this is in the top 25% of previous assessments in terms of reliability, and can be considered one of the best assessments in terms of reliability. If the value is, for example, 0.2, then this is in the bottom 25% of previous assessments in terms of reliability.

It is reasonable to say that Kappa values below 0.45 represent process assessments of the worst kind. This may be an indicator that the assessment method used is at fault. For example, previous research [11] has indicated that the assessment method affects reliability. Therefore, if the reliability is found to be "Poor", the assessors and the sponsor should consider changing the assessment method. A software process assessment should aim to attain at least a value of 0.63 or higher (i.e., the top 50% compared to previous assessments). The most reliable assessments will have values of 0.79 or higher.

This benchmark is based on *all* data available to date on the interrater agreement of process assessments, to our knowledge. This does not necessarily mean that it is representative of the reliability of *all* assessments. Therefore, when using the benchmark it is important to note that it provides an interpretation compared to the reliability of assessments conducted within the SPICE Trials thus far. As more data is collected, the benchmark can be adjusted accordingly.

In order to use the benchmark for evaluating the quality of an assessment, then certain precautions should be exercised for assessments utilizing more than one assessor. If within a particular organization an assessment is normally conducted with 2 assessors, then an interrater agreement evaluation ought to use kx2 person groups. This would provide results directly applicable to 2-assessor assessments. In such a case, ours can be considered to be a pessimistic benchmark since the data used to construct the benchmark comes from kx1 assessor assessments. This is based on the assumption that a 2 person team, for example, is more likely to produce repeatable ratings than a 1 person team, and a 3 person team would produce more repeatable ratings than a 2 person team, and so on. If this assumption is true, then the benchmark is most powerful at identifying low quality assessments which, from a practical perspective, is of most value.

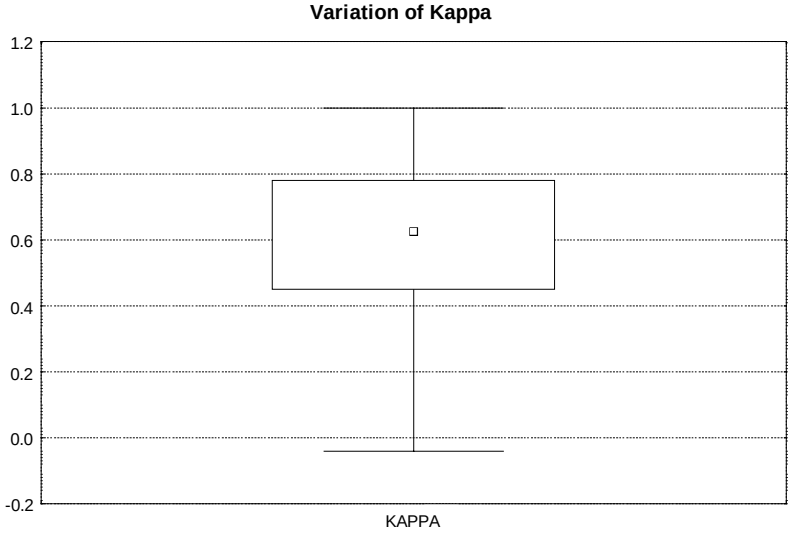

Figure 3: Box and whisker plot showing the distribution of Kappa.

Kappa Statistic	Strength of Agreement
<0.45	Poor (bottom 25%)
0.45-0.62	Moderate (bottom 50%)
0.63-0.78	Substantial (top 50%)
> 0.78	Excellent (top 25%)

Table 6: SPICE Software Process Assessment Kappa benchmark.

Interestingly, our benchmark has strong similarities to the Fleiss benchmark where Kappa values lower than 0.40 are considered to be poor, and values greater than 0.75 are considered to be excellent. Using our benchmark, Kappa values below 0.45 are considered to be in the worst class, and those above 0.78 are best in class.

Fleiss [16] provides essentially the same benchmark (as his unweighted Kappa benchmark) for a weighted version of Kappa that takes into account the seriousness of disagreement. This may suggest that the benchmark in Table 6 should also be applied to weighted versions of Kappa. However, the difficulty is that the values of Kappa will depend on the weighting scheme that is used. To date, no completely satisfactory weighting scheme for evaluating the disagreements in ISO/IEC 15504 assessments, or any process assessments in general, has been devised. Therefore, it would be prudent to devise an appropriate weighting scheme first and then construct a weighted Kappa benchmark for it.

5. Conclusions

Contemporary studies of the reliability of software process assessments frequently use the Kappa statistic of Cohen. However, the only existing benchmarks for interpreting the obtained values of Kappa (i.e., how good the reliability of an assessment is) came from the social science and medical literature. A priori this jump across disciplines makes the use of such benchmarks questionable. Therefore, in this paper a benchmark specific to software process assessments was constructed. The benchmark utilized data from the SPICE Trials. It can be used to decide the extent to which the reliability of new assessments is good or bad compared to assessments conducted within the SPICE Trials.

The impact that the reliability of assessments has on decisions made using assessment scores has not been addressed in this paper. This would be a basis for a different type of benchmark of the reliability of assessments, and should be a topic for future research. As for the current benchmark, the intention is to revise it as more data is collected from the SPICE Trials.

6. References

[1] M. Allen and W. Yen: *Introduction to Measurement Theory*. Brooks/Cole Publishing Company, 1979.

[2] D. Altman: *Practical Statistics for Medical Research*. Chapman and Hall, 1991.

[3] P. Armitage and G. Berry: *Statistical Methods in Medical Research*. Blackwell Science, 1994.

[4] L. Briand, K. El Emam, O. Laitenberger, and T. Fussbroich: "Using Simulation to Build Inspection Efficiency Benchmarks for Development Projects". To appear in *Proceedings of the International Conference on Software Engineering*, 1998.

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

[6] 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, 1995.

[7] 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.

[8] K. El Emam, L. Briand, and B. Smith: "Assessor Agreement in Rating SPICE Processes". In *Software Process Improvement and Practice Journal*, 2(4):291-306, 1996.

[9] 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.

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

[11] K. El Emam, B. Smith, and P. Fusaro: "Modeling the Reliability of SPICE Based Assessments". In *Proceedings of the Third IEEE International Software Engineering Standards Symposium*, pages 69-82, 1997.

[12] 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, 1998.

[13] K. El Emam and P. Marshall: "Interrater Agreement in Assessment Ratings". In K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press, 1998.

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

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

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

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

[18] P. Fusaro, K. El Emam, and B. Smith: "Evaluating the Interrater Agreement of Process Capability Ratings". In *Proceedings of the Fourth International Software Metrics Symposium*, pages 2-11, 1997.

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

[20] L. Gordis: *Epidemiology*. W. B. Saunders, 1996.

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

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

[23] H. Lyman: *Test Scores and What They Mean*. Prentice-Hall, 1963.

[24] F. Maclennan, G. Ostrolenk, and M. Tobin: “Introduction to the SPICE Trials”. 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, 1998.

[25] 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, 1998.

[26] J-M Simon, K. El Emam, S. Rousseau, E. Jacquet, and F. Babey: “The Reliability of ISO/IEC PDTR 15504 Assessments”. To appear in *Software Process Improvement and Practice Journal*.

[27] U. Umesh, R. Peterson, and M. Sauber: “Interjudge Agreement and the Maximum Value of Kappa”. In *Educational and Psychological Measurement*, 49:835-850, 1989.

[28] I. Woodman and R. Hunter: “Analysis of Assessment Data from Phase 1 of the SPICE Trials”. In *IEEE TCSE Software Process Newsletter*, No. 6, Spring 1996.

Document Information

Title:	Benchmarking Kappa for Software Process Assessment Reliability Studies
Date:	May 1998
Report:	IESE-016.98/E
Status:	Final
Distribution:	Public
	also published as ISERN-98-02