The Internal Consistency of Key Process Areas in the Capability Maturity Model® (CMM®) for Software (SW-CMM)

Ho-Won Jung
Dennis R. Goldenson

December 2002

TECHNICAL REPORT
CMU/SEI-2002-TR-037
ESC-TR-2002-037

Carnegie Mellon
Software Engineering Institute

Pittsburgh, PA 15213-3890

The Internal Consistency of Key Process Areas in the Capability Maturity Model® (CMM®) for Software (SW-CMM)

CMU/SEI-2002-TR-037
ESC-TR-2002-037

Ho-Won Jung
Dennis R. Goldenson

December 2002

Software Engineering Measurement and Analysis Initiative

This report was prepared for the

SEI Joint Program Office
HQ ESC/DIB
5 Eglin Street
Hanscom AFB, MA 01731-2116

FOR THE COMMANDER

Christos Scondras
Chief of Programs, XPK

This work is sponsored by the U.S. Department of Defense. The Software Engineering Institute is a federally funded research and development center sponsored by the U.S. Department of Defense.

Table of Contents

List of Figures

List of Tables

Acknowledgements

The authors wish to acknowledge the assessors, sponsors, and others who participate in assessments of the Capability Maturity Model® (CMM®) for Software (SW-CMM). This work would not be possible without the information that they regularly provide to the Software Engineering Institute (SEI℠). Thanks are due to Mike Zuccher, Kenny Smith, and Xiaobo Zhou for their support in extracting the data on which the study is based. The authors would also like to thank Sheila Rosenthal for her expert support with our bibliography, and Lauren Heinz for helping improve the readability of the document.

The authors express their thanks to our SEI colleagues, Will Hayes, Mike Konrad, Steve Masters, Jim McCurley, Mark Paulk, Mike Phillips, and Dave Zubrow. Thanks also to Khaled El-Emam, Robin Hunter, and Hyung-Min Park for their valuable comments on earlier drafts. We offer special thanks to Clifford E. Lunneborg, Professor Emeritus at the University of Washington, for providing helpful tips on S-Plus.

® Capability Maturity Model and CMM are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

℠ SEI is a service mark of Carnegie Mellon University.

Abstract

Evaluating the reliability of maturity level ratings is crucial for providing confidence in the results of software process assessments. This report examines the dimensions underlying the maturity construct in the Capability Maturity Model® (CMM®) for Software (SW-CMM) and then estimates the internal consistency (reliability) of each dimension. The analysis is based on 676 CMM-Based Appraisal for Internal Process Improvement (CBA IPI) assessments conducted during the period of January 2000 through April 2002. The results suggest that the SW-CMM maturity is a three-dimensional construct, with "Project Implementation" representing the maturity level 2 key process areas (KPAs), "Organization Implementation" representing the maturity level 3 KPAs, and "Quantitative Process Management" representing the KPAs at both maturity levels 4 and 5. The internal consistency for each of the three dimensions as estimated by Cronbach's alpha exceeds the recommended value of 0.9. Although more should be learned about the distinctions between maturity levels 4 and 5, the internal consistency of those KPAs is comparable to those at levels 2 and 3.

1 Introduction

The Capability Maturity Model® (CMM®) for Software (SW-CMM) is both a reference model for appraising software process maturity and a normative model for helping software organizations progress along an evolutionary path from ad hoc, chaotic processes to mature, disciplined software processes [Paulk et al. 93a-93c]. The CMM-Based Appraisal[1] for Internal Process Improvement (CBA IPI) is an assessment method to perform reliable and consistent assessments.

The SW-CMM is one of the best-known and most widely used models of its kind. Thousands of organizations have performed SW-CMM assessments and the resources expended on SW-CMM-based software process improvement (SPI) are estimated to be in the billions of dollars [Herbsleb et al. 97]. Hereafter this study interchangeably uses the two terms SW-CMM assessment and CBA IPI assessment unless there is reason to distinguish them.

Appraisal results have been used as a basis for many important decisions, including actions to improve internal software processes, large-scale acquisitions, and contract monitoring. For example, CMM level 3 has been encouraged for potential contractors of the U.S. Air Force [Saiedian & Kuzara 95, Coffman & Thompson 97]. Given the importance of the decisions influenced by appraisals and the resources required to implement them, both contractors and acquirers must be confident in the appraisal results.

1.1 The Reliability of Process Assessments

Increased confidence in assessment results can be achieved by demonstrating the reliability of assessment procedures. *Reliability* is defined as the extent to which the same measurement

® Capability Maturity Model and CMM are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

[1] *Appraisal* is a generic term used to refer to the diagnostic method independent of the context and motivation for its application. The term "appraisal" covers both assessment and evaluation. *Assessment* has come to connote the use of a diagnostic method for internal process improvement (e.g., self-examination) purposes. *Evaluation* has come to connote the use of a diagnostic method to provide insight to a separate (typically external) organization, frequently for purposes of acquisition or contract monitoring [Dunaway 96]. The purpose for its use may be different, but the diagnostic method itself is quite similar in both instances. Since our study uses a dataset from appraisals for internal process improvement, we intentionally use the term assessment.

procedure yields the same results on repeated trials [Carmines & Zeller 79]. Lack of reliability is caused by measurement error.

Similar to any other measurement procedure,[2] it is crucial to estimate the amount of measurement error in a SW-CMM assessment in order to provide confidence in the trustworthiness of its results. Measurement is defined as "the process of linking abstract *concepts* to empirical *indicants*" [Blalock 68]. The abstract concepts (theoretical constructs) are neither directly measurable nor observable, but can only be estimated by empirical indicants (indicators, items, measures, scale, or variables). In SW-CMM context, maturity is an abstract concept that is indirectly measured by using key process areas (KPAs).

So what would be the desirable qualities of measures? They are reliability and validity[3] [Carmines & Zeller 79]. Reliability concerns the degree of *repeatability* and *consistency* of empirical measurements [Zeller & Carmines 80]. The amount of random measurement error[4] is inversely related to the degree of reliability of the measuring instrument. Any measuring instrument is relatively reliable if it is minimally affected by random measurement error, where the term "instrument" implies a questionnaire, assessment procedure, or any other form of data collection that is used in rating software engineering practices. A set of KPAs is a type of instrument to collect data for measuring the maturity of organizations.

The more consistent the results given by repeated measurements, the higher the reliability of the measurement procedure. The consistency of the measurements is affected by ambiguities in wording and inconsistencies in interpretations by assessors [El-Emam & Goldenson 95, Fusaro et al. 98]. A survey of process assessments based on ISO/IEC 15504[5] shows that clarity of the semantics of the process definition in the 15504 document set is a third important variable[6] among 24 that affect reliability [El-Emam et al. 97]. Recent studies also show that more reliable assessments can reduce assessment effort during consolidation [El-Emam et al.

---

2 Assessors rate whether or not KPA goals are achieved. One or more goal ratings are combined to determine the level of KPA satisfaction. In turn, those measures of KPA satisfaction are combined to determine the capability maturity of a software organization. Thus, both KPA satisfaction and maturity levels are in fact derived measures [ISO 01].

3 *Validity* is defined as the extent to which any instrument measures what it is intended to measure. In other words, validity is related to accuracy, whereas *reliability* is related to repeatability and consistency. The notion of validity relates to "the assumption that measures of theoretic concepts should behave similarly toward theoretically relevant external variables" [Balch 74]. See Zeller and Carmines for general theory [Zeller & Carmines 80] and El-Emam and Birk for validity studies in software process assessments [El-Emam & Birk 00a-00b].

4 Systematic error does not affect reliability but affects validity. Since this study is limited to reliability, systematic error is not considered in this analysis.

5 ISO/IEC 15504 (*Software Process Assessment*) is a suite of international standards under development by Working Group 10 of Subcommittee 7 (Software Engineering Standardization) under Joint Technical Committee 1 for the International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) [ISO 98].

6 The two most important factors in that study are "Lead assessor's experience/competence in conducting assessment" and "Lead assessor's knowledge of ISO/IEC 15504 or WG10 documents."

98, Jung et al. 01]. In addition, questionnaire-based studies of SW-CMM assessment team leaders and team members indicate that the consolidation of data is one of the most difficult aspects of assessment [Dunaway & Baker 01].

Skepticism does remain in our field, both about the value of process improvement in general and the credibility of assessment results. Indeed, some critics have argued that little or no evidence exists [Fayad & Laitinen 97], while others have expressed concerns about the reliability of appraisal results [Bollinger & McGowan 91, Gray & Smith 98]. The critics are correct that credible evidence is vital. Other such evidence does exist, however, and it will be reviewed more fully later.

1.2 Study Purpose and Summary of Results

The objective of this report is to identify the dimensions underlying a set of SW-CMM KPA measures, and to then estimate the internal consistency (reliability) of each dimension of the capability maturity concept. Internal consistency is estimated using Cronbach's alpha [Cronbach 51], which is considered to be an appropriate method in the context of software process assessment and is commonly used in empirical software engineering. The results are based on maturity level data from 676 CBA IPI assessments conducted during the period from January 2000 through April 2002. To the authors' knowledge, this is the first study of its kind to investigate the dimensionality of organizational capability maturity and estimate internal consistency by using the results of full-scale SW-CMM assessments.

The report should serve to reduce concerns about reliability in SW-CMM assessments. Using factor analytic techniques, it is shown that the concept of capability maturity can be separated into three distinct dimensions. The maturity level 2 KPAs are in fact closely related as the SW-CMM model suggests, and they can be treated as a single "Project Implementation" dimension. Similarly, the maturity level 3 KPAs can be considered together as like aspects of "Organization Implementation." The maturity level 4 and 5 KPAs are combined under a single dimension termed "Quantitative Process Implementation." The values of Cronbach's alpha coefficient of internal consistency are higher than the recommended value of 0.9 [Fusaro et al. 98, Nunnally & Bernstein 94].

Section 2 provides a brief overview of the SW-CMM within the scope of the report and a brief theoretical background of the reliability theory focused on internal consistency in software process assessment. The section also describes previous studies of reliability in software process assessments. Section 3 addresses data collection, sampling characteristics, and analysis methods. The results of the analysis are presented in Section 4. Further discussion and conclusions are in Section 5.

2 Background and Literature Review

2.1 Appraisals of the CMM for Software

Process appraisals may be done for different reasons. CBA IPI assessments typically are intended to motivate organizations to initiate or continue software process improvement programs as well as to provide an accurate picture of the organization's maturity relative to the SW-CMM.[7] Software Capability Evaluations typically are used for source selection or contract monitoring. Although our results are based only on CBA IPI assessments, the issues of reliability that we discuss apply to all appraisals regardless of the purposes for which they are intended.

2.2 Estimating Measurement Reliability

There are a variety of reliability estimation methods, such as test-retest, alternative-form, split-half, and internal consistency[8] (Cronbach's alpha) [Zeller & Carmines 80]. This study uses Cronbach's alpha because it is known to be the most appropriate method for measuring reliability in software process assessments [El-Emam & Goldenson 00, Jung & Hunter 01]. It also is the most commonly used method in the software engineering community. Appendix B presents further theoretical concepts and empirical research methods for evaluating reliability in software process assessments.

2.2.1 Cronbach's Alpha

In discussing the reliability of measurements, a set of items (indicators) is posited to reflect an underlying construct. In the SW-CMM, maturity that is neither directly measurable nor observable can be indirectly measured by considering the assessed values of the KPAs. We can say that the SW-CMM uses an 18-item (or KPA) instrument to measure the maturity of

7 Those less familiar with CMM models and appraisal methods should see Appendix A for a fuller review of the SW-CMM, assessments, and maturity level determination.

8 In some studies, internal consistency includes average inter-item correlation, average item-total correlation, split-half, and Cronbach's alpha [Trochim 01]. However, this study uses internal consistency synonymously with Cronbach's alpha.

organizations. If the necessary data were readily available, we also could use a 52-item (goal) instrument.[9]

The type of scale used in most measurement instruments is a summative one [McIver & Carmines 81, Spector 92]. This means that the individual ratings x_i s for each item are summed up to produce an overall rating score, i.e., $y = \sum_{i=1}^{N} x_i$, where N is the number of items in an instrument. One property of the covariance matrix for a summative rating is that the sum of all terms in the matrix gives exactly the variance of the scale as a whole, i.e., $\sigma_y^2 = \sum_{(i,j)} \sigma_{ij}$, where σ_{ij} denotes covariance between items i and j ; if $i = j$, then $\sigma_{ij} = \sigma_i^2$.

The variability in a set of items score is considered to consist of two components as follows:

- The error terms are the source of unique variation that each item possesses, i.e. $\sum_{i=1}^{N} \sigma_i^2$.
- The signal component of variance that is considered to be attributable to a common source due to capability maturity is the difference between total variance and unique variance, i.e. $\sigma_y^2 - \sum_{i=1}^{N} \sigma_i^2$. Thus, the ratio of true to observed variance is $(\sigma_y^2 - \sum_{i=1}^{N} \sigma_i^2) / \sigma_y^2$.

To express this in relative terms, the number of elements in the covariance matrix of a summative rating must be considered. The total number of elements in covariance matrix is N^2, and the total number of communal elements is $N^2 - N$. Thus, Cronbach's alpha becomes:

$$\alpha = \frac{N}{(N-1)} \left[1 - \sum_{i=1}^{N} \sigma_i^2 \Big/ \sigma_y^2 \right] \text{ or } \alpha = \frac{N\bar{\rho}}{1 + \bar{\rho}(N-1)} ,$$

where N is the number of items; σ_i^2 and σ_y^2 are a unique variation of item i and total variation, respectively; $\bar{\rho}$ is equal to the mean inter-item correlation.

Cronbach's alpha is a generalization of Kuder-Richardson formula number 20 (KR20) to estimate the reliability of items scored dichotomously with zero or one [Kuder & Richardson 37]. KR20 is computed as follows:

$$\text{KR20} = \frac{N}{N-1} \left[1 - \sum_{i=1}^{N} p_i (1 - p_i) / \sigma_y^2 \right],$$

where N is the number of dichotomous items; p_i is the proportion responding "positively" to the item i ; σ_y^2 is equal to the variance of the total composite. KR20 has the same interpretation as Cronbach's alpha.

9 The SW-CMM includes 52 goals, with 20, 17, 6, and 9 goals in maturity levels 2, 3, 4, and 5, respectively.

Since KPAs in SW-CMM assessments are determined dichotomously as zero (“Not Satisfied”) or one (“Fully Satisfied”), this study can use KR20 without making any assumption about rating scale type. However, if a maturity or capability level is measured with a scale that uses more than two categories, such as in ISO/IEC 15504, reliability estimation does require assumptions about scale type. Since KR20 gives the same value as Cronbach’s alpha, this study uses the more popular term Cronbach’s alpha rather than KR20.

What constitutes a satisfactory Cronbach’s alpha value level of reliability depends on how a measure is being used. In the early stages of the research on an assessment instrument, reliabilities of 0.7 or higher are considered sufficient for narrow constructs [Cronbach 51, Nunnally & Bernstein 94] and 0.55 to 0.7 for moderately broad constructs [Van de Ven & Ferry 80]. For basic research, a value of 0.8 is acceptable. In applied settings where important decisions are being made with respect to assessment scores, a reliability of 0.9 is the minimum that would be acceptable [Nunnally & Bernstein 94].

Since maturity levels are in fact used in making important decisions, the minimum tolerable value of internal consistency in the SW-CMM should be set at 0.9. In ISO/IEC 15504, the minimum value also has been set at 0.9 [Fusaro et al. 98].

2.2.2 Dimensionality

Cronbach’s alpha assumes that the construct being measured is unidimensional [Carmines & Zeller 79]. As the name implies, unidimensional scaling is relevant to those situations in which it is presumed that there exists a single dimension underlying a set of data items [McIver & Carmines 81]. In contrast to unidimensional models, multidimensional scaling implies that there is more than a single dimension that underlies a set of items. If the SW-CMM maturity scale were multidimensional, then it would be more appropriate to compute the internal consistency for each dimension separately.

Two factor-analytic models, factor analysis and principle component analysis, can be used to investigate the dimensionality of process attributes. The objective of factor analysis is to search for, identify, or confirm the underlying factor(s) or construct(s) and to explain the correlation among items. The objective of principle component analysis is to reduce the number of variables to a few components, where each of the components can be represented as a linear combination of the corresponding variables [Sharma 96].

The scree plot [Cattell 66] and the eigenvalue-greater-than-one-rule[10] [Kaiser 70] are the two most popular methods to determine the number of factors (or components). The scree plot is a figure of the eigenvalues against the factor numbers. Cattell recommended retaining factors

---

[10] An eigenvalue represents the sum of squared factor loadings for all of the items in that factor. See Section 4.1.2 for an example.

above the "elbow" and rejecting those below it. The eigenvalue-greater-than-one-rule retains only the factors that have the eigenvalue of greater than 1. The rationale of this rule is that the amount of variance extracted by each factor should, at a minimum, be equal to the variance of at least one variable. Since the rule may lead to a greater or a fewer number of factors than are necessary [Cliff 88], it is recommended that this rule be used in conjunction with other rules [Sharma 96].

Factor loadings are the correlations between the items and the factors (underlying constructs). The loadings are criteria to determine the quality of factor classifications. Sharma recommended a cut-off value of 0.6 [Sharma 96]. Comrey provided a guideline of factor loading such as "fair" (0.45), "good" (more than 0.5), "very good" (0.63), and "excellent" (0.71) [Comrey 73].

2.3 Previous Studies of Reliability in Process Assessments

2.3.1 Internal Consistency

Most studies of reliability in software process assessments have been conducted as part of the Software Process Improvement and Capability dEtermination (SPICE) Trials[11] of the emerging International Standard ISO/IEC 15504. Two important objectives in the Trials were to evaluate reliability and validity of ISO/IEC 15504 conformant assessments [El-Emam & Goldenson 95]. The Trials team developed a study plan and concept of reliability in software process assessment based on ISO/IEC 15504. The Phase 2 SPICE Trials team published a summarized final result [Jung et al. 01].

Results from the Phase 2 SPICE Trials reported high internal consistency;[12] however, it became necessary to investigate any changes in reliability in the subsequent version of the Proposed Draft Technical Report (PDTR). Recently, Jung reevaluated the internal consistency of the ISO/IEC TR 15504 capability dimension [Jung 02a]. Results from assessments done in Korea again showed a high Cronbach's alpha value of 0.89 for capability levels 1 through 3, which is particularly noteworthy since the assessments were done using the English language version of the TR.

---

11 The SPICE Trials were performed in three broad phases. Phase 1 took place in 1995, Phase 2 from September 1996 to June 1998, and Phase 3 began in July 1998. In November 2001, SPICE Trials was reshaped with the name of SPICE Network. SPICE Network consists of SPICE Research, SPICE Benchmarking Forum, and SPICE Network Partner [ISO/WG10 01]. Empirical studies at Phase 2 were published [ISO/WG10 98 & 99].

12 Interrater agreement also has been used to estimate the reliability of software process appraisals. Estimates are made of the extent to which two assessors or teams of assessors agree when making independent judgments about the same software engineering processes. See Appendix B for further information.

The first reliability study of the 1987 maturity questionnaire was presented 10 years after its publication [Fusaro et al. 98]. Data for the study were from a Delphi panel and a mail survey. Fusaro et al. estimated an internal consistency of 0.94 by utilizing Cronbach's alpha, higher than the 0.90 reported earlier by Humphrey and Curtis. The internal consistency of the ISO/IEC PDTR[13] 15504 capability dimension was estimated at 0.90 in the same study. Moreover, Fusaro and his colleagues assumed that capability and maturity are unidimensional constructs. If they were multidimensional, their results would be deflated estimates.

Another SW-CMM study, based on a goal-level questionnaire administered to 45 projects in one large company, resulted in a Cronbach's alpha of 0.70 [Krishnan & Kellner 99]. However, they used a 5-point rating scale instead of the dichotomous scale typically used with the SW-CMM. Moreover they did not provide information about the dimensionality of their data. As just noted, Cronbach's alpha assumes unidimensionality. Their result is all the more compelling if there is in fact more than one separate dimension of capability maturity.

2.3.2 Dimensionality

Curtis collected questionnaire data on SW-CMM KPA goals covering maturity levels 2 and 3 from 3 organizations and performed principle component analysis for each organization [Curtis 96]. His results showed a multidimensional construct that included what he termed "planfulness," coordinated commitments, subcontractor management, quality assurance, configuration management, and process definition. Clark conducted a correlational study of levels 2 and 3 CMM KPAs using data collected from 50 organizations [Clark 97]. El-Emam and Goldenson performed a principal component analysis based on the Clark results and found a multidimensional factor structure that differed somewhat from Curtis' results [El-Emam & Goldenson 00].

The differences in the results of these studies arise from differences in their questions and data collection methods. The Curtis study is based on several individuals within the same organization. Clark's study is based on organizational-level data. Individual level design is useful for reliability studies, but the unit of observation ideally should be organization or project rather than individual [El-Emam & Goldenson 00]. In addition, Curtis examined KPA goal satisfaction, while Clark studied the implementation of KPAs. As will be seen in Section 4, both sets of results are different from our own, which are based on a much larger and varied dataset.

---

[13] ISO/IEC JTC1 has a variety of paths for developing International Standards [ISO 99]. One of them is through a published technical report (TR). A TR follows a series of stages such as NP (New Proposal), WD (Working Draft), PDTR (Proposed Draft Technical Report), DTR (Draft Technical Report), TR (Technical Report), and IS (International Standard). Assessments in the Phase 2 SPICE Trials were based on the PDTR version. At the time of writing, ISO/IEC 15504 was still in the second draft stage of TR (sometimes called TR2).

El-Emam investigated the dimensionality of the ISO/IEC 15504 capability scale [El-Emam 98]. He described two dimensions underlying a set of process attributes. The first dimension "Process Implementation," consists of the process attributes in capability levels 1 through 3. The second dimension, "Quantitative Process Implementation," covers the process attributes in capability levels 4 and 5. Jung and Hunter analyzed a total of 691 process instances assessed during the Phase 2 SPICE Trials (from September 1996 to June 1998) [Jung & Hunter 02]. They reconfirmed the multidimensionality of the ISO/IEC 15504 capability measures and provided Cronbach's alpha values of 0.88 and 0.87 for the two dimensions, respectively.

3 Research Method

3.1 Data Source

3.1.1 Data Collection

Lead assessors authorized by the Software Engineering Institute (SEI[SM]) are required to provide reports to the SEI for their completed assessments. Assessment data on the reports are kept in an SEI repository called the Process Appraisal Information System (PAIS).[14] The PAIS includes information for each assessment on company and appraised entity, KPA profiles, organization and project context, functional area representatives groups, findings, and related data.

Submitting an assessment report does not imply that the SEI certifies any assessment findings or maturity levels. All assessment data are kept confidential and are available only to SEI personnel on a need-to-know basis for research and development. Information in the PAIS is used to produce industry profiles or as aggregated data for research publications, and the SEI publishes a Maturity Profile report twice a year (http://pcaf/PAIS/).

The dataset that was analyzed for this study was extracted from appraisal reports in the PAIS for the period of January 2000 through April 2002. During the period, 948 appraisals were reported to the SEI, 707 (74.58%) of which were CBA IPI assessments of the SW-CMM. The remaining appraisals cover a variety of appraisal models and methods such as SW-CMM Software Capability Evaluation (SCE), CMM for Software Acquisition (SA-CMM), and CMM Integration[SM] (CMMI®).

Not all CBA IPI assessments include KPA rating profiles, since the determination of a maturity level or KPA ratings is optional and is provided at the discretion of the assessment sponsor. KPA ratings and maturity levels exist for 676 assessments, including 362 from organizations in the U.S. and 314 from non-U.S.-based organizations.

SM CMM Integration and SEI are service marks of Carnegie Mellon University.

14 The data entry form can be found in http://seir.sei.cmu.edu/ROE/.

® CMMI is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

Also note that Software Subcontract Management is excluded from the analyses, since that maturity level 2 KPA was not assessed by 74.56% (504/676) of the organizations in the sample.

3.1.2 Dataset Analyzed

Many assessments do not consider all KPAs up to and including maturity level 5. As seen in Table 1, four overlapping datasets were created based on the KPAs that were included in the model scope of the 676 assessments. By definition, all of the maturity level 2 KPAs were assessed in all 676 organizations. Those that assessed only the level 2 KPAs are removed from the second dataset, leaving 408 that assessed the KPAs through maturity level 3 or higher. Similarly, 156 assessments covered KPAs through level 4 or higher, and 78 covered the KPAs through level 5. Reading across the first row in Table 1, note that 93, 295, 176, 60, and 52 of the assessments were determined to be at maturity levels 1, 2, 3, 4, and 5, respectively. Reading down the columns, note also that 26 of the 93 organizations that were determined to be at maturity level 1 also were assessed against the level 3 KPAs. Two of them also were assessed against the level 4 KPAs, and one included the level 5 KPAs in the model scope of its assessment. Similarly, 94 of those determined to be at level 2 also were assessed unsuccessfully against some or all of the level 3, 4 or 5 KPAs. The same interpretation applies for the organizations that were determined to be at maturity levels 3 and 4. Of course, all of those who were determined to be at level 5 fully satisfied the goals of all of the KPAs.

Table 1: Number of Organizations at Each Maturity Level

Dataset	ML* 1	ML 2	ML 3	ML 4	ML 5	Total
Dataset 1: ML 2 KPAs' rating	93	295	176	60	52	676
Dataset 2: ML 2-3 KPAs' rating	26	94	176	60	52	408
Dataset 3: ML 2-4 KPAs' rating	1	9	34	60	52	156
Dataset 4: ML 2-5 KPAs' rating	1	7	10	8	52	78

* ML denotes maturity level.

We investigated the dimensions of maturity and then estimated internal consistency for each dimension separately for each of the four datasets. Consistent results from the four datasets provide increased confidence in our conclusions.

3.1.3 Unit of Analysis

In CBA IPI assessments, several projects are assessed in a single organization. The KPA profiles for an organization are the aggregate of assessment team judgments across those projects to produce a single maturity level for the entire organization within the scope of the assessment. Thus, the unit of analysis in this study is an organization.

Our dataset consists of KPA rating profiles from 676 organizations. In experimental terms, each of the 676 organizations becomes a case. Each case includes scores for that case on one or more attributes, where an *attribute* is defined as some characteristic of the *case* and the *score* is a value of the attribute. For instance, each assessment reported to the SEI becomes a single case, in which the case includes determinations (scores) of KPAs (attributes).

3.2 Sampling Characteristics of the Dataset

Statistical analyses and interpretations of the data at hand depend on the selection of a sample (subset) from a population. Population inference requires random sampling. Thus, we first examine the sampling characteristics of our dataset.

The simplest form of sampling is a (simple) *random sample*. The random sample is defined as “a set of cases selected from a well-defined population of cases by a process that ensures that every sample containing the same number of cases has the same chance of being the one selected” [Lunneborg 00]. In the SW-CMM assessment context, this definition explicitly implies two requirements: 1) a well-defined population of assessment cases from which to sample; and 2) a well-defined random process for selecting the sample.

The assessments reported to the PAIS database do not satisfy these two requirements. The population and the size of its assessments cannot be clearly defined, and the assessed organizations are not selected on a random basis. Rather, the assessments in PAIS are a self-selected sample, i.e., the assessed organizations voluntarily participated in CBA IPI assessments to improve their software process or were required to do so by the sponsors of their work. Hence, our analyses are based on nonrandom sampling methods.

In nonrandom design, the dataset itself in the PAIS is a population of assessment cases, where the population is called a *local population* or *a set of available cases* [Lunneborg 00]. In addition, a sample implies a random sample from that local population.

Because the cases are not a random sample, statistical inferences to a population beyond the dataset at hand are not possible. But, it is sensible to infer the *descriptions* to the local population. The descriptions are not inferences to a population. Rather, they should be considered as descriptive statistics, and they neither can be generalized to others nor have causal implications. Typical descriptions include measures of central tendency (e.g., means or medians), dispersion (e.g., variance or control limits), or relationship (e.g., correlation coefficients or internal consistency).

Descriptions based on a nonrandom sample need assurance that they truly characterize the available cases and that they are stable [Lunneborg 00, Montgomery et al. 98]. An available set of cases such as our assessment dataset cannot be assumed to have the same degree of homogeneity as a random sample. A fair description is a stable one that is relatively uninfluenced by the presence of specific cases. Thus, results of this report should be tested for their stability (homogeneity).

3.3 Data Analysis

As we just noted, since our dataset is not a random sample, we need to examine the stability of its estimated internal consistency. For this purpose, we use a subsample technique to evaluate the stability. We recompute internal consistency on a sequence of subsamples, where each subsample contains some but not all the available cases [Lunneborg 00]. If the internal consistency has apparently changed due to leaving out some cases, the reliability is not stable. The same statistical analyses usually can be used in both random and nonrandom cases, but a nonrandom case requires an assumption of stability.

3.3.1 Generating a Set of Bootstrap Subsamples

How one forms the subsamples depends on the information that is available about the manner in which the case data are collected, i.e., *structured* and *unstructured* datasets. In a structured dataset, the subsamples are formed based on contextual information such as time of data collection, site, or investigator (experimenter, teacher, etc). In contrast, sufficient contextual information is missing in an unstructured dataset. The structure of our dataset is unclear, except for the organizational location classifications of U.S. and non-U.S., but the usage of organizational location as a structure criterion results in an insufficient number of observations for estimating the alpha value at maturity level 5. Fortunately, however, resampling of an unstructured dataset also has the advantage of providing confidence intervals for Cronbach's alpha values.

The stability of internal consistency in an unstructured dataset is examined using a set of subsamples that are generated from a bootstrap[15] resampling procedure [Lunneborg 00].[16] The procedure draws a sample of size n without replacement, where n is the number of observations in the original dataset. Then, a subsample can be obtained by taking a half of the sample, i.e., a half-size sample. This process is repeated B times, where B is a large number as high as 1,000. The reason for choosing half-samples is that there will be a larger number of distinct subsamples of that size than of any other size [Lunneborg 00] that could be drawn from the full sample n. This study uses an S-Plus bootstrap routine to generate its half-size samples [Mathsoft 99].

15 This bootstrap method should not be confused with the Bootstrap model for process assessment [Kuvaja 99].

16 Evaluating stability in a structured dataset requires computing a description of the original dataset and then recomputing the same description for each subsample partitioned by time, site, or other criteria. If the internal consistency values of the subsamples are close to each other and to the description of the original dataset, then the description is considered to be stable. On the other hand, if the description of a particular subsample is remarkably different from those of the other subsamples and the original data set, this implies that the description of the original dataset depends heavily on the cases omitted from that subsample.

A bootstrap method has been successfully used previously in empirical software engineering. El-Emam and Garro estimated the number of SPICE assessments by utilizing a capture-recapture method [El-Emam & Garro 00]. Jung and Hunter utilized a bootstrap method in computing confidence levels for the capability levels for each ISO/IEC 15504 process [Jung & Hunter 01].

3.3.2 Examining the Stability of the Estimated Results

The bootstrap resampling procedure in this report can be used for parameter estimation or confidence intervals on the mean or median (difference). If 1,000 subsamples are taken, and an internal consistency value is computed for each of the 1,000 subsamples, then the lower and upper limits of the confidence interval for internal consistency can be determined at percentiles of 2.5% and 97.5% respectively. The histogram of 1,000 replications is called the *empirical reference distribution*. The confidence interval of the empirical reference distribution is called the *empirical confidence interval* (ECI).

Internal consistency from the original dataset should be solidly in the middle of the empirical reference distribution in order to be considered stable. It should not be at or near the limits of the empirical reference distribution. The difference between a value of internal consistency in the original dataset and the mean of those in B subsamples is called bias. For defining bias, define: t_b^* as a value of internal consistency at the b th subsample, where $b=1, \ldots, B$; B is the number of replications (here, 1000); $\hat{\theta}$ is an estimated internal consistency from the original dataset. Then the bias, $BIAS$, is defined as follows:

$$BIAS = \frac{\sum_{b=1}^{B} t_b^*}{B} - \hat{\theta} .$$

In addition, the sample-to-sample variability of the estimated internal consistency is the standard deviation of the sampling distribution of B replicates. This is called the standard error (SE) of the estimate and is defined as follows:

$$SE = \sqrt{\frac{\sum_{b=1}^{B}\left(t_b^* - \overline{t}^*\right)}{B-1}} \text{, where } \overline{t}^* = \sum_{b=1}^{B} t_b^* \Big/ B$$

The degree of bias is evaluated against the SE of the sampling distribution of B replicates. If the bias is large relative to the SE, there is a problem (unstable). A criterion for judgment is that if the absolute value of the bias is less than one-quarter the size of the SE, the bias can be ignored safely [Efron & Tibshirani 93]. One can conclude that the internal consistency from the original dataset is stable.

4 Results

4.1 Descriptive Summary of the Dataset

Our dataset is based on the 676 CBA IPI assessments that were reported to the SEI during the period of January 2000 through April 2002. Figure 1 shows the number of assessed organizations by site type in both the U.S. and non-U.S.[17] Over a third of the U.S. organizations are government contractors, but almost 60 percent produce products for the commercial market or for their own use in-house. A noticeable minority are themselves U.S. government and military organizations. While over 10 percent of the non-U.S. organizations are contractors for the U.S. government, the same proportion develop or maintain software for their own in-house use, and three-quarters of them develop or maintain software for commercial sales.

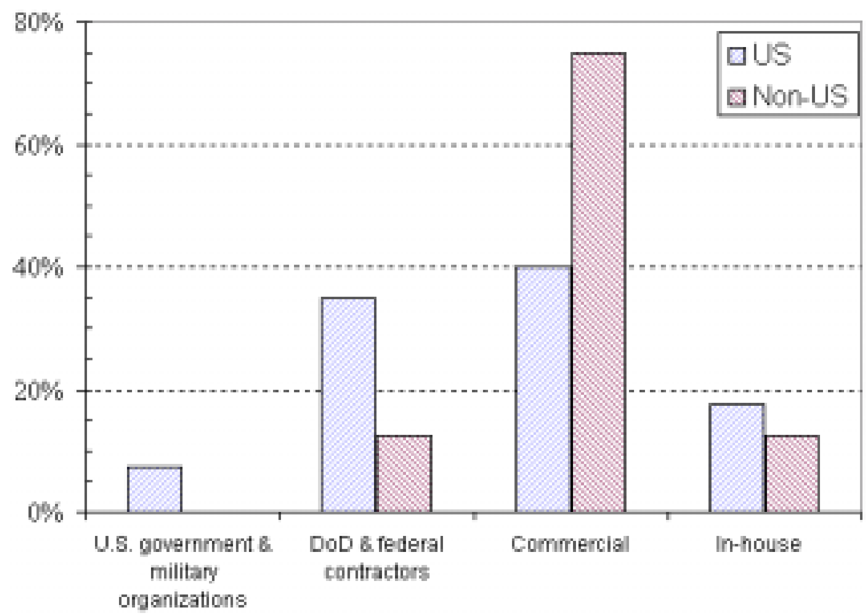

Figure 1: Types of Organization Assessed

The dataset includes 2,860 projects from 676 CBA IPI assessments. Figure 2 is a box and whisker plot showing the variation in the number of projects in the assessments (missing for 5 assessments).[18] The minimum and maximum numbers of projects accessed in one assess-

17 Two organizations are missing in each region.

18 See Appendix C for an explanation of box and whisker plots.

ment are 1 and 21, respectively. The mean and median projects assessed in a single assessment are 4.26 and 4, respectively. This number is essentially the same as the recommended 4 projects for CBA IPI assessments [Dunaway & Baker 01].

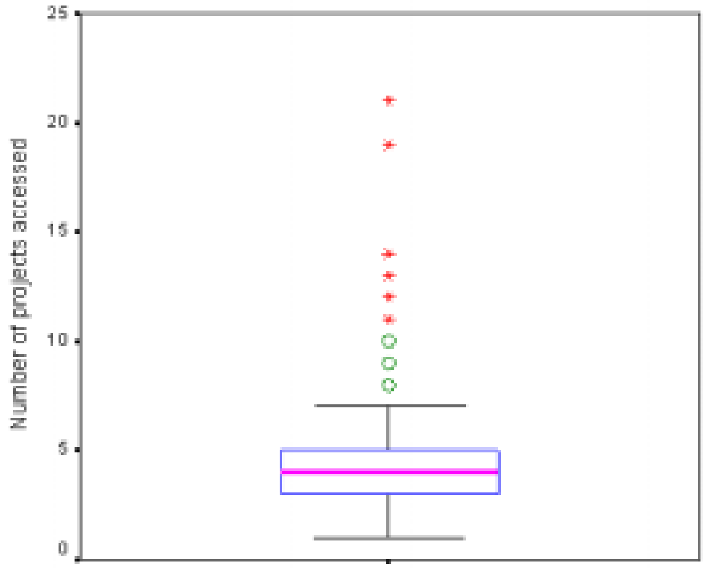

Figure 2: A Box and Whisker Plot Showing the Variation of Projects Assessed

Figure 3 shows the number of Functional Area Representative (FAR[19]) group interviews in 671 assessments. The mean is 6.06 and the median is 6. The maximum number of FAR groups interviewed is 18.

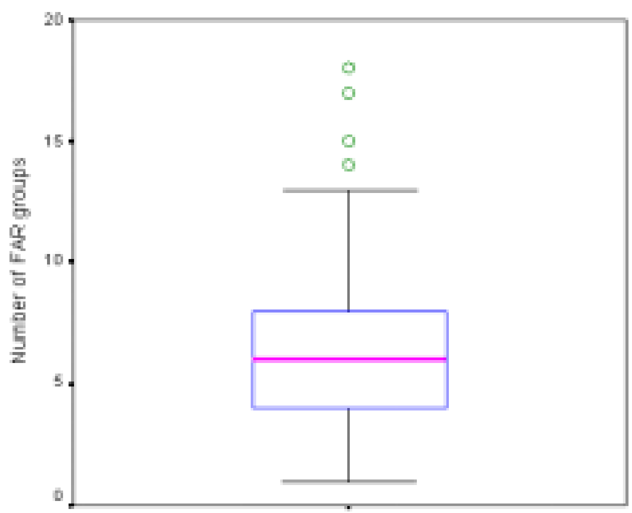

Figure 3: Number of FAR Groups Interviewed per Assessment

---

19 Functional Area Representatives are practitioners who have technical responsibilities in various areas that support their organizations' software development or maintenance projects, e.g., configuration management or quality assurance. Selected FAR interviewees should be a representative sample of the assessed organization's technical staff. FAR interviewees should be practitioners, not managers or staff. No two individuals who have a reporting relationship to each other should be in a FAR interview session together.

Figure 4 shows the total number of functional area representatives interviewed in the same 671 assessments. The mean and median are 32.50 and 30 respectively. In one case, there were 127 interviewees.

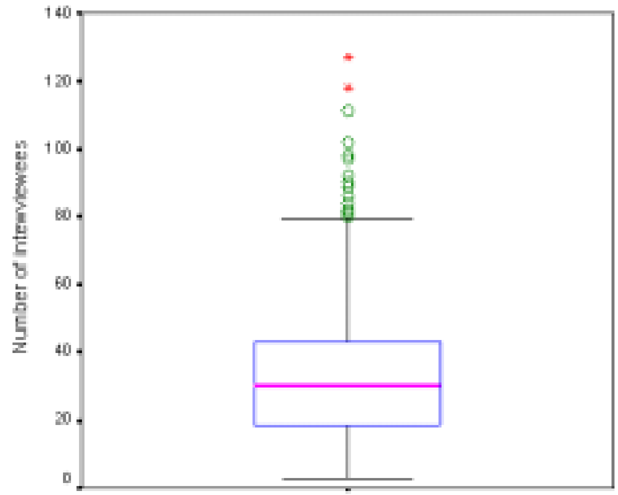

Figure 4: Number of Interviewees per Assessment

Figure 5 shows the average number of FAR group members per interview based on 670 assessments. The mean and median are 5.05 and 5.45, respectively. This number is within the recommended range of “four to eight participants” [Dunaway & Masters 96]. However, the number of organizations within the recommended range is 392 organizations (58.51%). Approximately 40% are out of the recommended range. As such, further investigation is required.

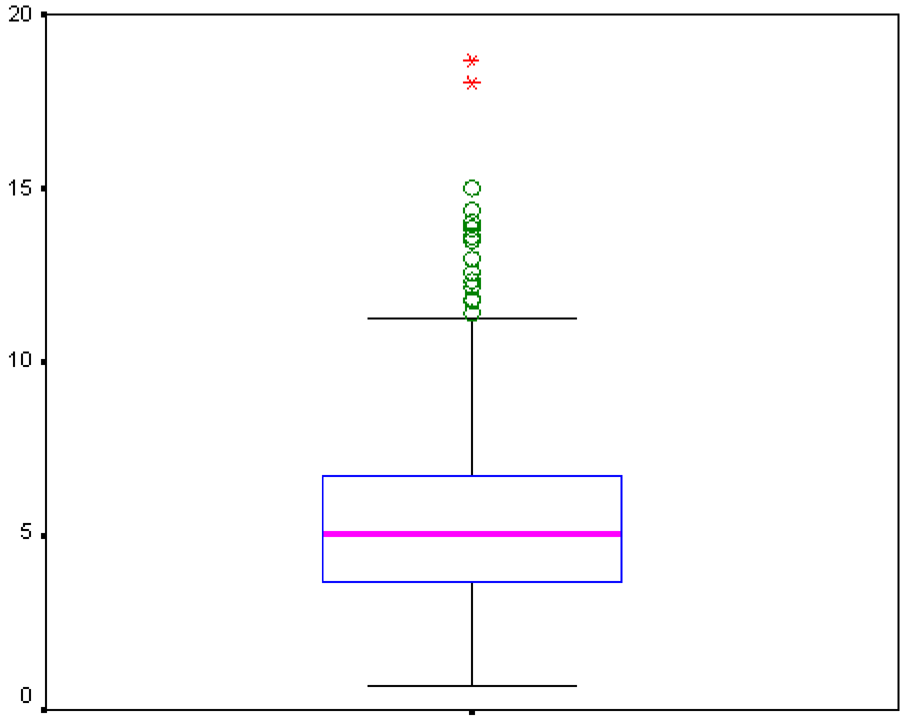

Figure 5: Average Number of FAR Interviewees

Figure 6 shows the distribution of the maturity level of the 676 organizations assessed. The most frequent level is 2 (Repeatable) with 43.64 % (295) of the organizations. The next was level 3 (Defined). The number of level 4 organizations was slightly higher than that of level 5, i.e. 8.88% (60) versus 7.69% (52).

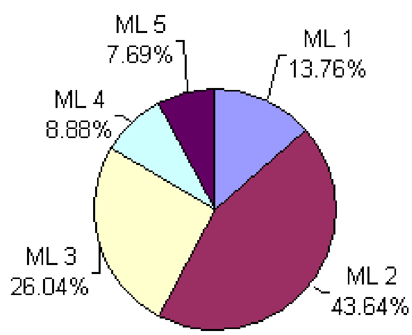

Figure 6: Distribution of Maturity Levels

It is hard to imagine that the proportion of organizations at maturity level 2 is larger than that at maturity level 1 in software organizations throughout the world. A study by Fayad and Laitnen indicates that most development organizations are at maturity level 1 [Fayad & Laitnen 97].

The fact that the dataset is not a random sample makes it impossible to interpret the distribution. There is no way to determine whether organizations assessed according to the SW-CMM are a representative (random sample) of the industry at large. It is most likely that, as early adopters of new technology, and specifically as organizations interested in SPI, these organizations are from the “high end” of the maturity spectrum. This phenomenon has been detected in the SPICE Trials as well [Rout et al. 98]. These results partially support the assumption of nonrandom sampling in this study.

4.2 Analysis Results

Analysis results are presented in the sequence of datasets summarized in Table 1. Cronbach’s alpha as a measure of internal consistency is computed separately for more than one dimension. Hence, we first describe the dimensions that underlie the KPAs in the SW-CMM.

4.2.1 Cronbach’s Alpha in Maturity Level 2 KPAs

As seen in Table 1, all of the five KPAs at maturity level 2 except Software Subcontract Management were rated in 676 organizations. To investigate the dimensionality, factor analysis with a principal component analysis was performed for the five KPAs. The scree plot of the five KPAs in Figure 7 shows a break after the first component (i.e., a unidimensional characteristic of the five KPAs). In addition, only one eigenvalue of 4.18 satisfies the greater-than-one rule. The scree plot and eigenvalue-greater-than-one rule reach the same conclusion. The factor loadings in Table 2 exceed the recommended cut-off value of 0.6 [Sharma 96] as well as an “excellent” criterion of 0.71 [Comrey 73]. Thus, we conclude that the five KPAs at maturity level 2 are items of a single construct, which we call the “Project Implementation” dimension.

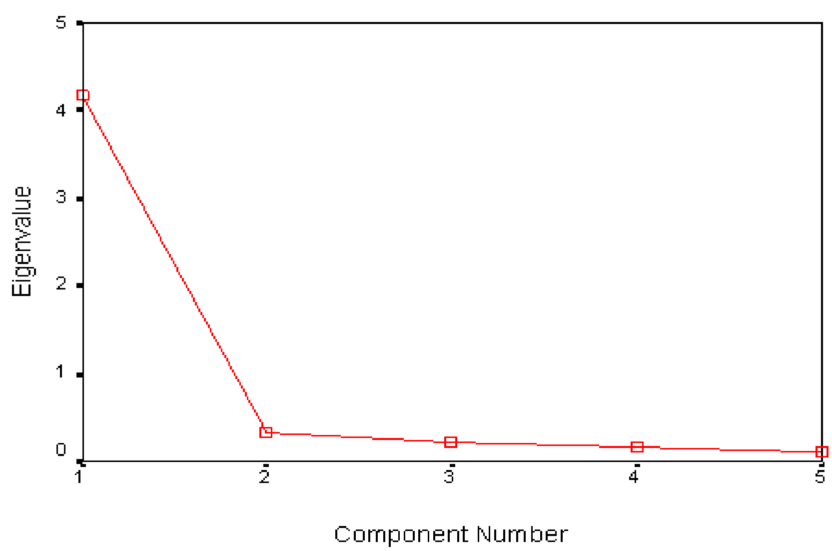

Figure 7: The Scree Plot of the Five KPAs at Maturity Level 2

A factor loading (λ_i) indicates the correlation between a variable (KPA determination) and the underlying factor. The eigenvalue equals the sum of the squared factor loadings for a factor over all items, i.e., $\sum \lambda_i^2 = 0.87^2 + 0.94^2 + 0.92^2 + 0.92^2 + 0.92^2$ =4.18. The value of 83.59%, in the last row of Table 2, denotes the proportion of variance explained by the common factor, i.e., the degree of factorial determination of variables. Its computation is performed by the eigenvalue divided by the number of items (i.e., 4.18/5=0.8359).

Table 2: Factor Loadings ( ) for the Five KPAs at Maturity Level 2

KPA name	**Factor 1 ("Project Implementation")**
Requirements Management	**0.87**
Software Project Planning	**0.94**
Software Project Tracking and Oversight	**0.92**
Software Quality Assurance	**0.92**
Software Configuration Management	**0.92**
% of variance explained	83.59%

The (observed) Cronbach's alpha of the five KPAs is 0.9495 as seen in Table 3. This value exceeds a standard recommendation of 0.9 [Fusaro et al. 98, Nunnally & Bernstein 94] and is high enough to use in practice.

Table 3: Cronbach's Alpha and Bootstrap Results in KPAs at Maturity Level 2

	Observed Cronbach's alpha	**Bootstrap Cronbach's alpha**	**Bias**	**SE**	**95% ECI**
Dimension 1 ("Project Implementation")	0. 9495	0.9499	0.0004	0.0077	[0.934, 0.964]

Figure 8 shows the distribution of the bootstrap mean of Cronbach's alpha in the maturity level 2 KPAs. The dotted and solid vertical lines in Figure 8 denote a bootstrap value of 20.9499 and observed value[20] of 0.9495, respectively. The absolute difference 0.0004 in alpha values between observed and bootstrap values is small relative to the SE value of 0.0077. This satisfies the criterion for evaluating stability [Efron & Tibshirani 93]. The ECI of 95% is [0.934, 0.964]. Thus, we can conclude that the resulting alpha value in KPAs of maturity level 2 is fairly stable. The lower limit of the CI is greater than 0.9 of a recommended value.

[20] The term *observed* implies "without bootstrap resampling."

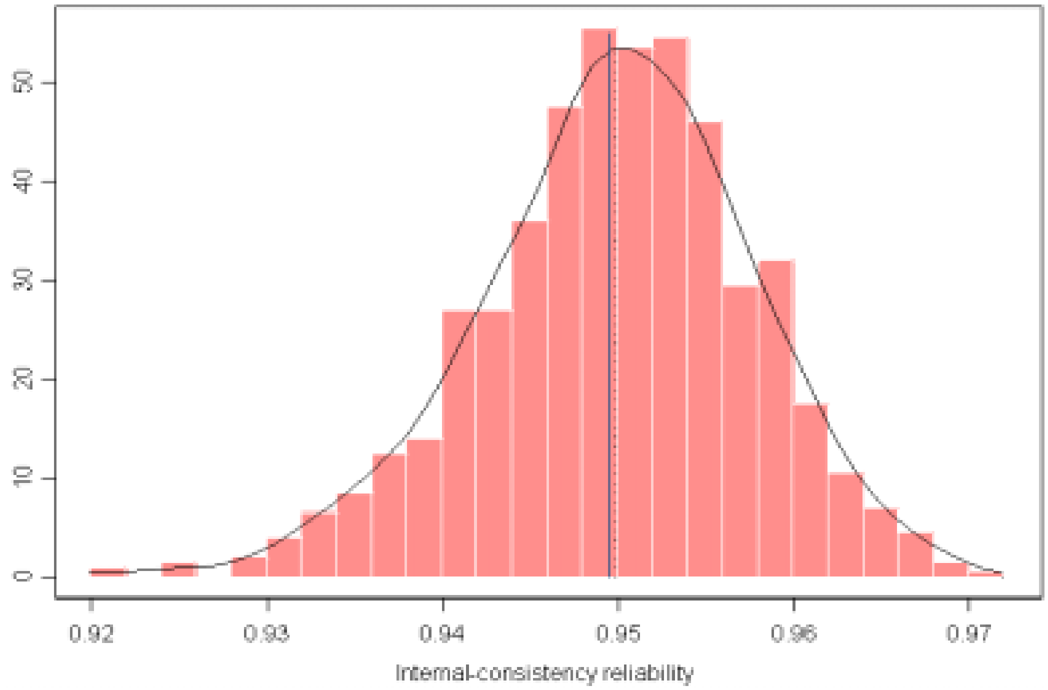

Figure 8:   Cronbach's Alpha in Maturity Level 2-3 KPAs

As seen in Table 1, 408 assessments covered 12 of the 13 KPAs in maturity levels 2 and 3 (i.e., 5 KPAs at level 2 and 7 KPAs at level 3). Of those organizations, 120 (29.41%) were assessed at level 1 or 2. The remaining 288 organizations were assessed at maturity level 3, 4, or 5.

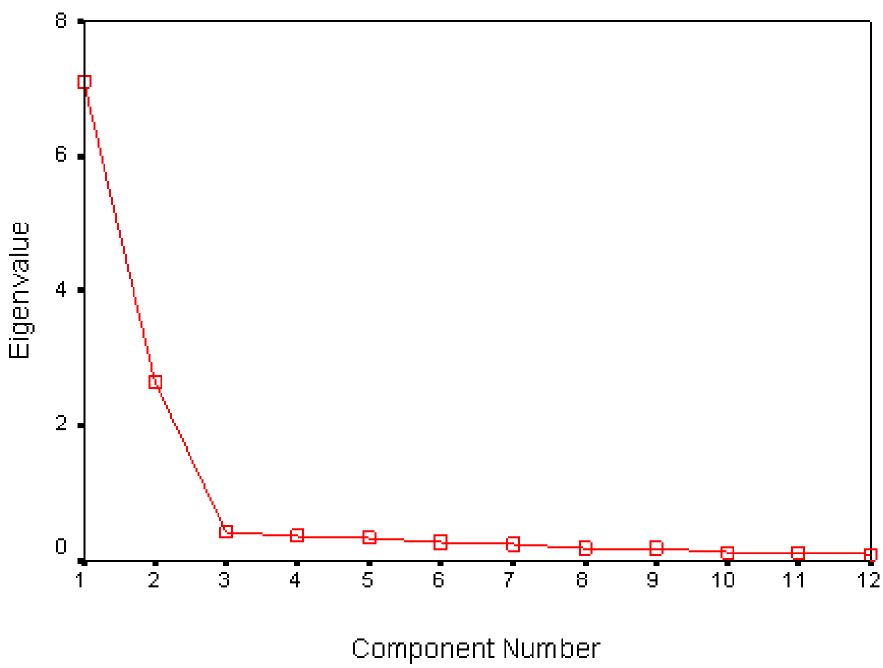

Figure 9:   The Scree Plot of the 12 KPAs at Maturity Levels 2-3

The scree plot of the 12 KPAs in Figure 9 shows a two-dimensional structure, where the eigenvalues greater than 1 are 7.11 and 2.65, respectively. Their factor loadings in Table 4 show that the KPAs at maturity levels 2 and 3 form two separate dimensions. This suggests that the 12 KPAs are in fact fairly well grouped and defined in the alignment postulated by the SW-CMM maturity level definitions. We call the second factor "Organization Implementation," which corresponds to the KPAs at maturity level 3.

This result is consistent with the one from dataset 1, which covers the KPAs at maturity level 2 only. The proportion of variance explained by the two factors is 81.27%, i.e., 22.06% + 59.21%.

Table 4: Factor Loadings for the KPAs at Maturity Levels 2-3

KPAs at maturity levels 2-3	**Factor 1 ("Project Implementation")**	**Factor 2 ("Organizational Implementation")**
Requirements Management (2)	**0.85**	0.17
Software Project Planning (2)	**0.92**	0.18
Software Project Tracking and Oversight (2)	**0.89**	0.19
Software Quality Assurance (2)	**0.86**	0.25
Software Configuration Management (2)	**0.88**	0.22
Organization Process Focus (3)	0.32	**0.78**
Organization Process Definition (3)	0.18	**0.92**
Training Program (3)	0.21	**0.88**
Integrated Software Management (3)	0.17	**0.92**
Software Product Engineering (3)	0.19	**0.89**
Intergroup Coordination (3)	0.17	**0.85**
Peer Review (3)	0.20	**0.87**
% of variance explained	22.06%	59.21%

* The number in parenthesis of KPA denotes maturity level.

Table 5 shows observed and bootstrap alpha values and 95% confidence intervals for each of the two dimensions. The alpha values are greater than the recommended value of 0.9. Comparisons of the bias and the SE values show stability of the two alpha values. The bootstrap distribution (not shown) also corroborates the stability of the alpha values.

Table 5: Cronbach's Alpha and Bootstrap Results in KPAs at Maturity Levels 2-3

	Observed Cronbach's alpha	**Bootstrap Cronbach's alpha**	**Bias**	**SE**	**95% ECI**
Dimension 1 ("Project Implementation")	0.942	0.939	-0.003	0.019	[0.895, 0.969]
Dimension 2 ("Organization Implementation")	0.960	0.959	-0.001	0.005	[0.950, 0.968]

4.2.2 Cronbach's Alpha in Maturity Level 2-4 KPAs

As shown in Table 1, 14 KPAs at maturity levels 2 through 4 (i.e., 5 KPAs at level 2, 7 KPAs at level 3, and 2 KPAs at level 4) were rated in 156 organizations. Among them, 44 organizations did not attain maturity level 4. The remaining 112 organizations were determined to be at maturity level 4 or 5.

The scree plot of the 14 KPAs in Figure 10 shows a break after the first three components, with eigenvalues of 8.061, 3.657, and 1.543, respectively. The scree plot and eigenvalue-greater-than-one rule suggests three-dimensional structure of the maturity concept based on these KPAs. Table 6 demonstrates high factor loadings for each of the three dimensions, which in fact correspond to the structure postulated by the SW-CMM for these KPAs. We conclude that each stage of maturity levels 2 through 4 corresponds to a separate dimension of the maturity concept. We call the third dimension "Quantitative Process Implementation."

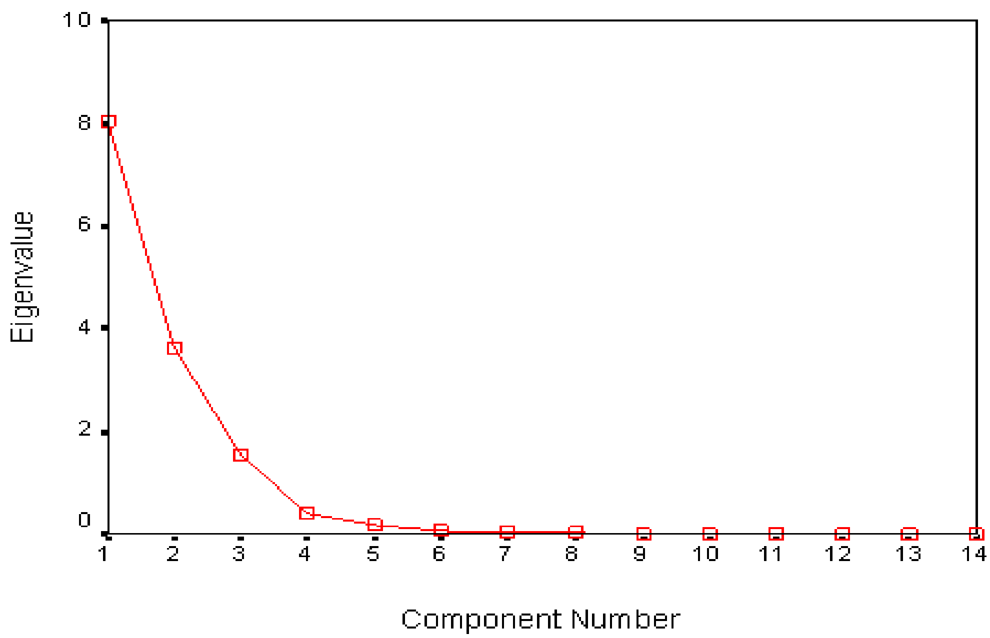

Figure 10: The Scree Plot of the 14 KPAs at Maturity Levels 2-4

Cronbach's alpha and bootstrap values are shown in Table 7. The alpha values for each of the three dimensions are 1, 0.981, and 0.968, which are very acceptable levels. The values of bias and SE indicate stability of the three alpha values.

Table 6: Factor Loading for the KPAs at Maturity Levels 2-4

KPAs in maturity levels 2-4	Factor 1 ("Project Implementation")	Factor 2 ("Organizational Implementation")	Factor 3 ("Quantitative Process Implementation")
Requirements Management (2)	**0.98**	0.18	0.04
Software Project Planning (2)	**0.98**	0.18	0.04
Software Project Tracking and Oversight (2)	**0.98**	0.18	0.04
Software Quality Assurance (2)	**0.98**	0.18	0.04
Software Configuration Management (2)	**0.98**	0.18	0.04
Organization Process Focus (3)	0.20	**0.93**	0.12
Organization Process Definition (3)	0.15	**0.94**	0.14
Training Program (3)	0.18	**0.90**	0.15
Integrated Software Management (3)	0.14	**0.90**	0.18
Software Product Engineering (3)	0.16	**0.92**	0.15
Intergroup Coordination (3)	0.15	**0.94**	0.14
Peer Review (3)	0.20	**0.93**	0.12
Quantitative Process Management (4)	0.04	0.25	**0.95**
Software Quality Management (4)	0.05	0.22	**0.96**
% of variance explained	26.12%	57.58%	11.02%

Table 7: Cronbach's Alpha and Bootstrap Results in KPAs at Maturity Levels 2-4

	Observed Cronbach's alpha	Bootstrap Cronbach's alpha	Bias	SE	95% ECI
Dimension 1 ("Project Implementation")	1	1	-	-	-
Dimension 2 ("Organization Implementation")	0.981	0.977	-0.004	0.039	[0.924, 1]
Dimension 3 ("Quantitative Process Implementation")	0.968	0.967	-0.001	0.017	[0.911, 1]

4.2.3 Cronbach's Alpha in Maturity Level 2-5 KPAs

A total of 78 organizations were assessed through maturity level 5. Among them, 54 organizations were determined to be at level 5. The scree plot of the 17 KPAs in Figure 11 shows a three-dimensional structure of the maturity concept, with eigenvalues greater than 1 of 9.94, 4.07, and 2.31, respectively. The factor loadings in Table 8 show that the first two dimensions remain the same as in our previous results; however, the KPAs at maturity levels 4 and 5 form a single dimension. We have called the third dimension "Quantitative Process Implementation." At least in this dataset, the KPA profiles across maturity levels 4 and 5 are in fact closely interrelated.

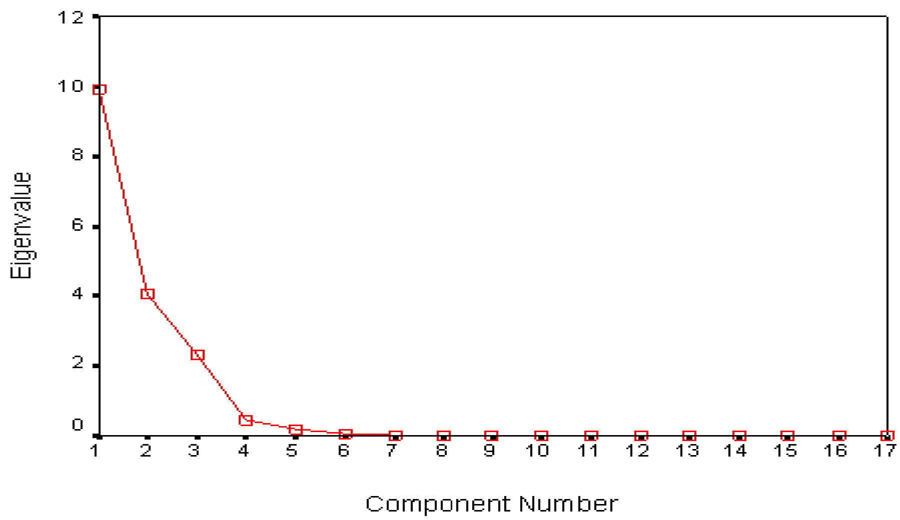

Figure 11: The Scree Plot of the 17 KPAs at Maturity Levels 2-5

Table 8: Factor Loading for KPAs at Maturity Levels 2-5

KPAs	**Factor 1 (“Project Implementation”)**	**Factor 2 (“Organizational Implementation”)**	**Factor 3 (“Quantitative Process Implementation”)**
Requirements Management (2)	**0.98**	0.18	0.07
Software Project Planning (2)	**0.98**	0.18	0.07
Software Project Tracking and Oversight (2)	**0.98**	0.18	0.07
Software Quality Assurance (2)	**0.98**	0.18	0.07
Software Configuration Management (2)	**0.98**	0.18	0.07
Organization Process Focus (3)	0.18	**0.95**	0.24
Organization Process Definition (3)	0.18	**0.95**	0.24
Training Program (3)	0.16	**0.90**	0.31
Integrated Software Management (3)	0.18	**0.95**	0.24
Software Product Engineering (3)	0.16	**0.90**	0.31
Intergroup Coordination (3)	0.18	**0.95**	0.24
Peer Review (3)	0.18	**0.95**	0.24
Quantitative Process Management (4)	0.08	0.39	**0.84**
Software Quality Management (4)	0.08	0.39	**0.84**
Defect Prevention (5)	0.06	0.21	**0.93**
Technology Change Management (5)	0.06	0.22	**0.94**
Process Change Management (5)	0.06	0.24	**0.94**
% of variance explained	23.93%	58.44%	13.59%

Once again, the Cronbach’s alpha values of 1, 0.995 and 0.970 are at very acceptable levels. As shown in Table 9, bootstrap results verify the stability of the three alpha values.

Table 9: Cronbach's Alpha and Bootstrap Results in KPAs at Maturity Levels 2-5

	Observed Cronbach's alpha	**Bootstrap Cronbach's alpha**	**Bias**	**SE**	**95% ECI**
Dimension 1 ("Project Implementation")	1	1			
Dimension 2 ("Organization Implementation")	0. 995	0. 993	-0.002	0. 023	[0.983, 1]
Dimension 3 ("Quantitative Process Implementation")	0. 970	0. 970	0	0. 011	[0.947, 0. 989]

5 Discussion and Conclusions

5.1 Interpreting the Results

First and foremost, our results provide confidence in the internal consistency of the KPAs of the SW-CMM. The consistently high values of Cronbach's alpha show that CBA IPI assessment teams typically do in fact make rating judgments that are internally consistent with the structure of the SW-CMM.

Our factor analytic results also lend credibility to the structure of the SW-CMM itself. As the model posits, there is independent empirical evidence that affinity groupings of KPAs should be considered as separate dimensions of the capability maturity construct. Moreover, the "Project Implementation" and "Organization Implementation" dimensions are entirely congruent with the KPAs at maturity levels 2 and 3, respectively.

More does need to be learned, however, about the distinctions between levels 4 and 5. The maturity level 4 and 5 KPAs form a single dimension that we have termed "Quantitative Process Implementation." The KPA profiles across maturity levels 4 and 5 are in fact closely interrelated in the dataset we analyzed. The KPAs at maturity levels 4 and 5 all necessitate statistical thinking and the careful use of quantitative analytic methods, so it does make sense that their satisfaction profiles are closely interrelated. Still, our present results are based on a relatively small number of high maturity organizations, and one certainly could argue that it remains best practice to master the level 4 KPAs first. Clearly, more and better evidence is needed to provide a fuller understanding of the nuances in the results, along with opportunities for improvement in the model and appraisal methods.

Recall from our literature review in Section 2 that the concept of ISO/IEC 15504 capability consists of a two dimensional structure [El-Emam 98, Jung & Hunter 02]. Our own results support the validity of separate dimensions corresponding to maturity levels 2 and 3 respectively. The existence of an additional dimension may reflect the difference between organizational maturity as characterized by the SW-CMM and process capability as characterized by ISO/IEC 15504. Note too that both the 15504 studies and our own suggest the existence a similar "Quantitative Process Implementation" construct. Such similarities based on disparate methods and data provide additional confidence about both sets of results.

Regardless, the dimensionality that we found holds consistently for four datasets, and internal consistency as estimated by Cronbach's alpha is consistently high for each of the three dimensions. Although the KPAs at levels 4 and 5 map to the same dimension in our current analysis, we can safely conclude that each stage of the SW-CMM corresponds to a dimension of organizational maturity and that the KPAs in each maturity level are good items for measuring the maturity of a common underlying construct.

5.2 Methodological Issues

5.2.1 Rating Scales

It is well known that the choice of rating scale affects estimates of internal consistency. If too few categories are used, the rating scale does not capture the full discriminatory power of the measures. On the other hand, using too many categories may be beyond the limited discriminatory powers of assessors.

In a reliability study of attitude scales with 3, 5, and 7 categories, Likert and Roslow concluded that the five-point scales consistently produced higher reliability estimates than did the others.[21] Similarly, a Monte Carlo study of the effects on reliability of the number of scale points showed that reliability estimates increased as the number of scale points increased from two to five, but the estimates decreased as more categories were added [Lissitz & Green 75]. Another similar study reached the same conclusions [Van de Ven & Ferry 80]. Finally, Jung and Hunter showed that the current four-point scale to rate ISO/IEC 15504 process attributes cannot be improved in terms of internal consistency by reducing it to 3 or 2 categories [Jung & Hunter 02].

Assessments based on the SW-CMM typically have used a two-point scale of "Fully Satisfied" and "Not Satisfied" to measure the extent of achievement of a KPA goal. Krishnan and Kellner addressed the difficulty in using a two-category scale when there is inconsistency in the extent to which SW-CMM activities and practices are implemented and institutionalized [Krishnan & Kellner 99]. Based on an earlier proposal, they used a five-point scale, since two categories cannot measure partial achievement.[22] The Standard CMMI Appraisal Method for Process Improvement (SCAMPI[SM]) now uses a very similar four-point scale to characterize practice implementation [SEI 01]. We next propose to replicate and extend this study for CMMI-based appraisals.

---

21 Likert, R. & Roslow, S. *The Effects upon the Reliability of Attitude Scales of Using Three, Five, Seven Alternatives*. Working paper, New York University, 1934.

22 This is taken from a 1994 presentation by Dennis Goldenson titled "A Multiple Response Scale for Process Measurement."

SM SCAMPI is a service mark of Carnegie Mellon University.

5.2.2 Explaining the High Internal Consistency

How can we explain the high Cronbach's alpha value over 0.9 for each of the three dimensions of capability maturity? The likeliest explanation is that our data come from well-trained assessors using a standardized assessment procedure (CBA IPI) against a well understood reference model (the SW-CMM). The importance of both training and a standardized method is described elsewhere [El-Emam & Madhavji 95, Trochim 01].

Though it is difficult to directly compare the Cronbach's alpha values, those reported here are consistently higher than those found in similar previous studies. It seems natural, though, that data drawn from actual assessments would be more internally consistent than those from a Delphi panel, questionnaires, or a mail survey. A full-scale assessment is based on several sources of objective evidence, which gives the appraisal team the opportunity to clarify any apparent ambiguities.

5.3 Future Research

Although the SEI's PAIS database retains the largest number of assessment cases available anywhere, the dataset is not a random sample, and our results cannot be generalized to all SW-CMM assessments conducted around the world. This sometimes is referred to as a threat to external validity, due to the type of sampling [Trochim 01]. Hence interpretation of our results should rightly be limited to CBA IPI assessments reported to PAIS by the current base of CMM users. However, if any other SW-CMM assessments also satisfy the same requirements specified by CBA IPI, then it can be expected that they too will be internally consistent with our results.

No single study can be fully definitive. Similar studies must be conducted that include assessment results that may not be well represented currently in the PAIS database. Such studies should include sample surveys as well as results from mini assessments and similar "lighter weight" appraisals conducted on organizations that are not yet ready to invest in a full, comprehensive appraisal.

A generalization of internal consistency is not easy work. "Even if a high level of internal consistency is achieved in the initial item analysis, it is a good idea to replicate it in subsequent samples. Availability of reliability estimates across different types of samples will expand the generalization of the scales' reliability…. There should be little variation in the magnitude of the internal consistency from sample to sample" [Spector 92, p. 65]. Changes over time and perturbations in the ways SW-CMM assessments are conducted around the world require careful monitoring of internal consistency. Replications are necessary to raise the confidence in any findings.

Additional empirical work also is necessary to help us better understand the underlying dimensional structure of capability maturity. This study will be replicated as soon as sufficient data have been reported for CMMI appraisals. Possibly using data from other sources, patterns of anomalies in the satisfaction of goals out of the order prescribed by CMM models should also be more closely examined.

The high internal consistency demonstrated among related KPAs suggests another question worthy of further study. Namely, is it possible to achieve comparable levels of reliability by sampling from within each dimension? Of course, organizations that are concerned about a particular KPA must consider the evidence in full detail. Is it possible, though, to make sufficiently confident statements about maturity levels for some purposes by examining evidence from a subset of model practices?

Appendix A A Brief Review of the CMM for Software and Appraisal Methods

The SW-CMM

The SW-CMM, on which process assessments that we have studied are based, codifies what many experts believe to be best practices of software engineering. The SW-CMM was the first in what has become a family of Capability Maturity Models. These include most notably the Software Acquisition Capability Maturity Model (SA-CMM) [Cooper & Fisher 02], the Systems Engineering Capability Maturity Model (SE-CMM) [Bate et al. 95], the People Capability Maturity Model (P-CMM) [Curtis et al. 01], and the Integrated Product Development Capability Maturity Model (IPD-CMM). Much of the content of these models now is being superseded by the suite of CMM Integration (CMMI) models [SEI 02].

Early versions of Humphrey's maturity framework appear in technical reports [Humphrey 87, Humphrey & Sweet 87], in papers [Humphrey 88], and in his book, *Managing the Software Process* [Humphrey 89]. A preliminary maturity questionnaire [Humphrey & Sweet 87] was released in 1987 as a tool to provide organizations with a way to characterize the maturity of their software processes. The questionnaire was not recommended as a stand-alone tool for estimating the maturity of an organization. Rather, it was meant as an orientation tool to be used during an assessment [Olson et al. 1989]. All of these efforts were initiated in response to a request to provide the federal government with a method for assessing the capability of its software contractors.

After four years of experience with the software process maturity framework and the preliminary version of the maturity questionnaire, the SEI and its affiliates evolved the software process maturity framework into a fully defined model: SW-CMM Version 1.0. Two technical reports, *Capability Maturity Model for Software* [Paulk et al. 91] and *Key Practices for the Capability Maturity Model for Software* [Weber et al. 91] formalized the description of the maturity levels in terms of KPAs. With the result of feedback from the software community, Version 1.0 was evolved to Version 1.1 as two technical reports, *Capability Maturity Model for Software, Version 1.1* [Paulk et al. 93a] and *Key Practices for the Capability Maturity Model for Software* [Paulk et al. 93b]. A detailed history can be found in the SEI publications [Paulk et al. 94, Paulk 95].

The SW-CMM provides a framework for organizing software processes into five evolutionary steps, or maturity levels, which lay successive foundations for continuous process improvement (see Table 10 [Paulk 99]). The SW-CMM covers practices for planning, engineering, and managing software development and maintenance. When followed, these key practices are meant to improve the ability of organizations to meet goals for cost, schedule, functionality, product quality, and other performance objectives.

Table 10: Maturity Levels and their Key Process Areas

Level	Focus	Key Process Areas
Level 5 Optimizing	Continuous process improvement	- Defect Prevention - Technology Change Management - Process Change Management
Level 4 Managed	Product and process quality	- Quantitative Process Management - Software Quality Management
Level 3 Defined	Engineering processes and organizational support	- Organization Process Focus - Organization Process Definition - Training Program - Integrated Software Management - Software Product Engineering - Intergroup Coordination - Peer Review
Level 2 Repeatable	Project management processes	- Requirements Management - Software Project Planning - Software Project Tracking and Oversight - Software Subcontract Management - Software Quality Assurance - Software Configuration Management
Level 1 Initial	Competent people (and heroics)	

Table 10 and Figure 12 show the structure of the Capability Maturity Model. As shown in Table 10, with the exception of Level 1, each maturity level is composed of several KPAs that indicate the areas on which an organization should focus to improve its software process. Each KPA is organized into five common features, including Commitment to Perform, Ability to Perform, Activities Performed, Measurement and Analysis, and Verifying Implementation. These attributes serve to indicate whether the implementation and institutionalization of a KPA is likely to be effective, repeatable, and lasting. The common features specify the key practices that, when collectively addressed, are meant to accomplish the goals of the KPA. The key practices describe activities and infrastructure that are believed to contribute most to the effective implementation and institutionalization of the KPA. The SW-CMM maturity levels denote stages[23] along an evolutionary path of software process improvement. Thus, the

[23] Software process assessment models in which only certain processes are assessed at each (maturity) level are referred to as staged models whereas software process assessment models in which each process may be assessed at each (capability) level are referred to as continuous models. Thus the SW-CMM is a staged model and ISO/IEC 15504 is a framework for continuous models. However, some processes in ISO/IEC 15504 are strongly related to particular capability levels, so the distinction between continuous and staged models may not be as clear as it may at first appear.

maturity level is an ordinal scale for measuring the maturity of an organization's software process and for evaluating its software process capability.

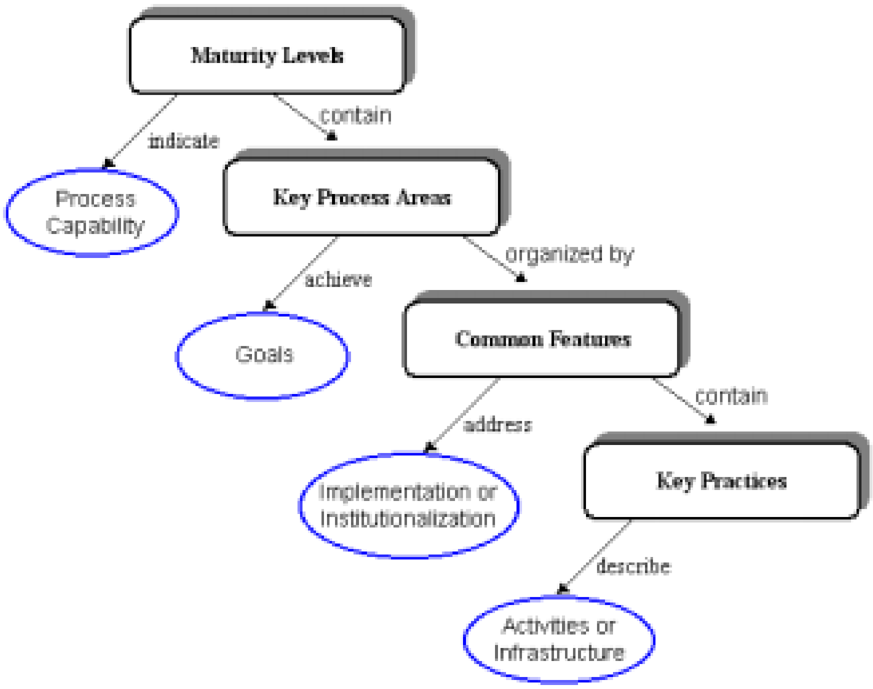

Figure 12: The Structure of the Capability Maturity Model

Appraisal Methods

An appraisal method describes the activities and procedures that need to be conducted during an assessment. It also includes the identification of the assessment sponsor, the purpose and scope of the assessment, any relevant constraints, assessment responsibilities, and so on.

Several appraisal methods have been developed since the first SEI-assisted assessment was conducted. A method for process assessment, called Software Process Assessment (SPA), was commercialized to industry and government licensees in 1990. SW-CMM Version 1.0 and Version 1.1 were published in 1991 and 1993, respectively. Since the SPA preceded the SW-CMM publication, the SEI developed the CBA IPI method for assessing an organization's software process capability. CBA IPI Version 1.0 was released in 1995 and updated to CBA IPI Version 1.1 in 1996 [Dunaway 96, Dunaway & Masters 96]. CBA IPI officially replaced SPA in 1995 through expiration of all SPA licenses effective on 31 December 1995. In 2002, the SEI released Version 1.2 [Dunaway & Masters 01].

The similarities and differences between the two models are described by Paulk [Paulk 99].
CMMI models each have two representations, both continuous and staged.

After an assessment that is done for purposes of internal process improvement, the senior manager of the assessed organization retains the assessment findings and results, and generally uses them to formulate an action plan for the process improvement program. Analysis of assessment results in light of an organization's business needs can identify the strengths, weaknesses, and risks inherent in that organization's processes. This, in turn, often leads to an improvement initiative that aims to make the processes more effective in achieving their goals, and to prevent significant causes of poor quality, cost overruns, or schedule delay.

CBA IPI Version 1.1 specifies minimum requirements for the composition of an assessment team, an assessment plan, data collection, data validation, rating, and reporting the results. CBA IPI also defines three phases of an assessment. The first phase includes the activities necessary to plan and prepare for the assessment (see Figure 13).The second phase consists of on-site activities for conducting the assessment, including techniques for gathering, organizing, and consolidating data (see Figure 14). The final phase is to report the results. Each phase is described in CBA IPI Version 1.1 in detail.

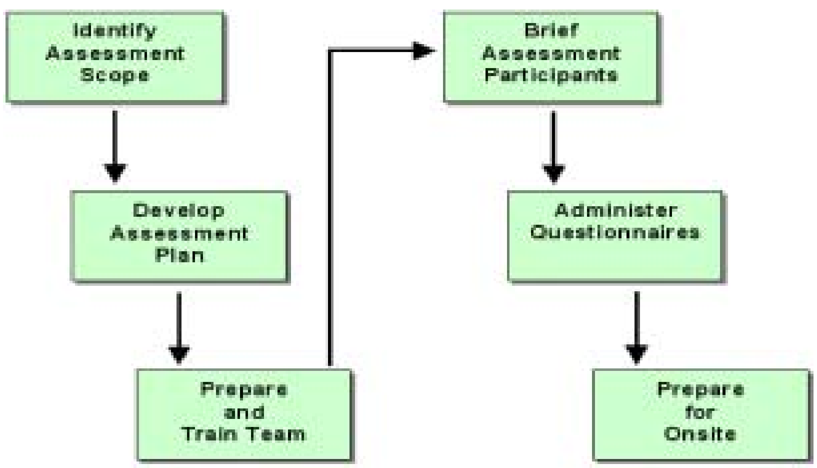

Figure 13: Pre-Onsite Activities

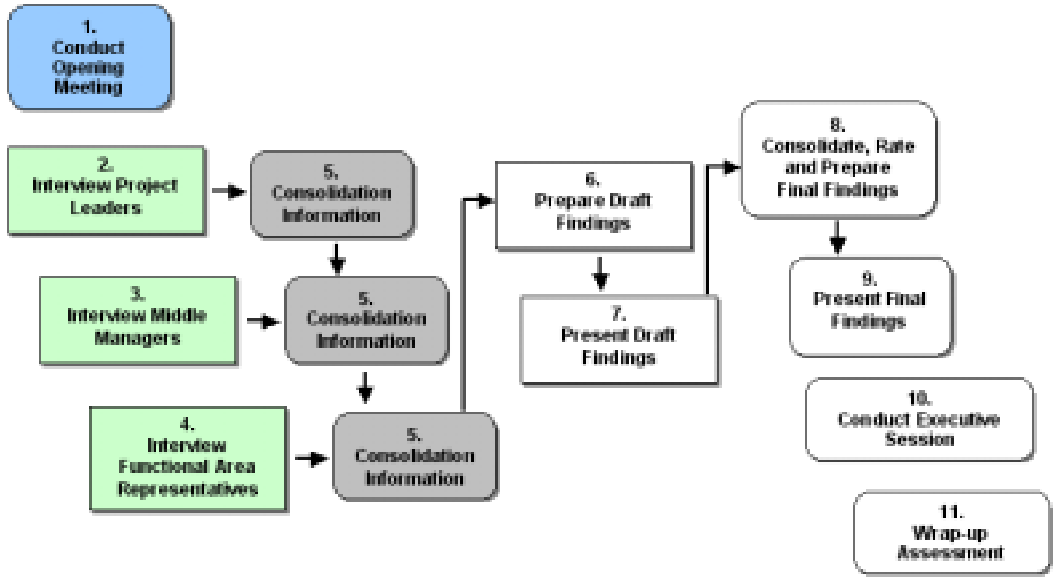

Figure 14: Chronology of On-Site Activities

Software Capability Evaluation (SCE) is an appraisal method that is meant for use in source selection and contract monitoring [Byrnes & Phillips 96]. SCE offers a means to help acquisition managers to identify program risk by evaluating software process capability in source selection and manage program risk by motivating contractors to improve their software development processes without forcing compliance to specific practices. Sometimes, SCEs are used to guide internal process improvement.

In 1995, the SEI published the CMM Appraisal Framework (CAF) [Masters & Bothwell 95]. The CAF describes the common requirements used for developing appraisal methods based on SW-CMM Version 1.1 and also provides a basis for comparing assessment and evaluation results. Both the CBA IPI and SCE methods are CAF-compliant.

Maturity Level Determination

A maturity level is defined in terms of satisfaction of the goals of the KPAs within its level. Maturity levels 2-5 each have several KPAs associated with them, as shown in Table 10. Each KPA is rated as either “Fully Satisfied” or “Not Satisfied,” and the maturity level is determined by aggregating the KPA ratings.

The rating scheme in CMM assessments allows the aggregation of judgments across several projects to produce a single maturity level for the entire organization within the scope of the assessment. An organization is defined to be at maturity level k if all KPAs up to and including maturity level k are rated as “Fully Satisfied.” For example, to determine whether an organization has achieved maturity level 2 or not, it is necessary to determine the ratings achieved by the six KPAs defined at maturity level 2. An organization that fails to achieve maturity level 2 is determined to be at maturity level 1.

Appendix B Reliability in Measuring Capability Maturity[24]

Reliability estimation in measurement has been based on the classical test (score) theory [Carmines & Zeller 79]. The theory is represented by an observed rating X with two additive components of a true rating T and an error term, i.e., $X = T + e$, which is a special case in which there is no systematic error. Since the measurement errors e are random, the observed ratings are sometimes higher or lower than the true rating. Therefore, in the long run, the mean of the error terms becomes zero. The *reliability of measurement* is defined as the ratio of true to observed variance. If no random error is involved in the measurement (i.e., the variance of the random term; $\mathrm{var}(e) = 0$), then the reliability equals 1. If the observed variance is equal to random variance (i.e. $\mathrm{var}(T) = 0$), it implies zero reliability in the measurement. Since the variance of the true rating, $\mathrm{var}(T)$, cannot be measured, it is estimated.

Reliability estimation methods such as test-retest, alternative-form, split-half, and internal consistency (Cronbach's alpha) can be categorized into *stability* (repeatability) and *equivalence* (consistency) by the basic strategies used to evaluate reliability [Zeller & Carmines 80].

Stability (repeatability) implies that repeated assessments of the same process, by the same assessor using the same or an alternative instrument, at two different points in time, should produce results that can be accepted as being identical [Zeller & Carmines 80]. It is estimated by utilizing the test-retest method or the alternative-form method. Consistency (equivalence) in a measurement focuses on multiple items of a concept measured at a single point in time, where each item is considered a separate but equivalent measure of the underlying concept. The split-half and internal consistency methods are used to measure consistency.

The first two methods (test-retest and alternative-form) for repeatability have major limitations in process assessments, such as requiring two consequent assessments by the same assessors with the same instrument (in test-retest) or two different instruments (in alternative-form). The split-half method has difficulties in dividing items and can be considered as a specific case of Cronbach's alpha. The four methods are explained in the context of software process assessments in detail elsewhere [El-Emam & Goldenson 00]. El-Emam & Goldenson consider internal consistency to be the most appropriate method for measuring reliability in software process assessments.

---

[24] This section is partially based on materials from Jung [Jung 2002a].

Since there is more than one method for estimating reliability, a method must be selected. The selection should consider the advantages and disadvantages of each method in a specific application field, as well as perceptions in the research community about what are appropriate methods of reliability estimation. Research communities related to software processes are management information systems (MIS) and software engineering. In both communities, internal consistency is the most popular method for estimating reliability of measurement.

Cronbach's alpha has been used for many years to estimate the reliability of MIS measurement instruments. Examples include software processes and their outcomes [Subramanian & Nilakanta 94], systems effectiveness [Srinivasan 85], user information satisfaction [Ives et al. 83, Tait & Vessey 88], user involvement [Amoako-Gyampah & White 93, Baroudi et al. 86], perceived ease of use and usefulness of software [Adams et al. 92, Davis 89], information system service quality [Jiang et al. 00, Van Dyke et al. 99], evaluation of information systems [Goodhue 98], organizational benefits of IS projects [Mirani & Lederer 98], and satisfaction with high-speed networks [Eum et al. 2001]. Sethi and King note that Cronbach's alpha is the most important method for estimating reliability of instruments [Sethi & King 91]. Test-retests have been used to measure reliability of a user information satisfaction instrument [Galletta & Lederer 89] and of a user involvement instrument [Torkzadeh & Doll 94].

Internal consistency is the most popular method to estimate reliability in software engineering as well. For example, Cronbach's alpha was used for estimating the internal consistency of the 1987 maturity questionnaire and the ISO/IEC 15504 capability dimension [Fusaro et al. 98], of an organizational maturity instrument [El-Emam & Madhavji 95], of key success factors in SPI [Dybå 00, El-Emam et al. 01], and of the ISO/IEC PDTR 15504 capability dimension [El-Emam 98, Jung & Hunter 02].

Interrater agreement also has been used to estimate the reliability of software process appraisals. Estimates are made of the extent to which two assessors or teams of assessors agree when making independent judgments about the same software engineering processes. Internal consistency and interrater agreement are sometimes called internal reliability and external reliability, respectively [Fusaro et al. 98].

In practice, the subjective nature of ratings makes it unlikely that there will be perfect interrater agreement; however, a series of such studies conducted as part of the international SPICE trials does show reasonably high levels of interrater agreement [El-Emam 99, El-Emam & Goldenson 00, Jung, et al. 01]. More recent work by Jung provides further discussion about paradoxes in the interpretation of the Kappa coefficient that is used in the SPICE studies.[25]

---

[25] Described in an article by Ho-Won Jung, titled "Evaluating the Interrater Agreement in SPICE-based Software Process Assessments," which was under review at the time of this publication.

Additionally, results from a CMMI pilot appraisal show remarkably similar results from two parallel teams that independently prepared observations, findings, and goal ratings for the entire scope of a maturity level 5 appraisal.[26] For example, only 5 of 79 goal ratings differed, all of which were due to difficulty interpreting new model content. These studies and other studies also have begun to analyze the (validity) factors that appear to affect differences in interrater reliability.

[26] This material is taken from a presentation by Dennis Goldenson, titled “But Can I Trust the Appraisal Results? Existing Evidence and What’s Coming Next,” delivered to the National Research Council, Canada, in October 2002.

Appendix C Explanation of a Box and Whisker Plot

The box and whisker plot below (Figure 15) provides a graphical presentation of data for displaying various features such as dispersion, location, and skewness. The bottom of the box corresponds to the first quartile (Q_1) and indicates the value of the variable to which 25% of the observations are less than or equal. Similarly the top of the box corresponds to the third quartile (Q_3). The length of the box called the IQR (interquartile range) is a measure of the dispersion of the data. A line within the box indicates the median (the 50th percentile), which is the statistic that indicates the center of the distribution. The median line in this study is drawn with a bold line to avoid an overlay of both end lines of the box and the medium. Two whiskers are extended from the box. The lower whisker starts at $\max\{X_{(1)},\ Q_1 - 1.5(Q_3 - Q_1)\}$ and the upper whisker ends at $\min\{X_{(n)},\ Q_1 + 1.5(Q_3 - Q_1)\}$, where $X_{(1)}$ and $X_{(n)}$ are the smallest and largest value of observations. Outliers are data points beyond the lower and upper whiskers, and they are plotted with **o**'s. Extreme values are data points beyond the outliers, and they are plotted with asterisks *****.

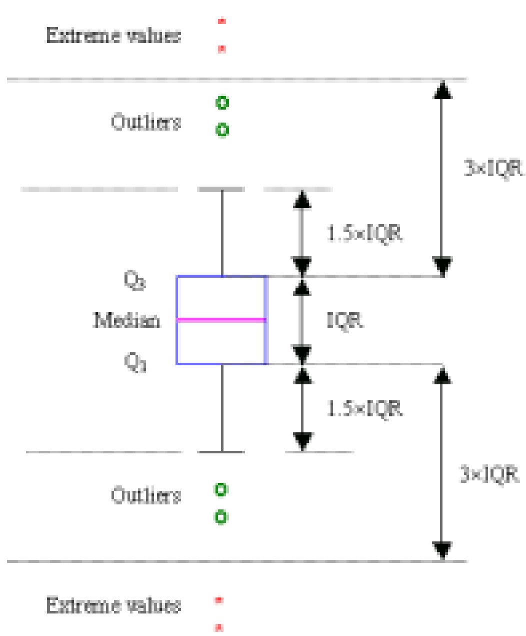

Figure 15: Explanation of a Box and Whisker Plot

References/Bibliography

[Adams et al. 92] Adams, D.; Nelson, R.; & Todd, P. “Perceived Usefulness, Ease of Use, and Usage of Information Technology: A Replication.” *Management Information Systems Quarterly 16,* 2 (June 1992): 227-247.

[Amoko-Gyampath & White 93] Amoako-Gyampah, K. & White, K. “User Involvement and User Satisfaction: An Exploratory Contingency Model.” *Information and Management 25,* 1 (July 1993): 1-10.

[Balch 74] Balch, G. I. “Multiple Indicators in Survey Research: The Concept ‘Sense of Political Efficacy.’” *Political Methodology 1*, 2 (Spring 1974): 1-43.

[Baroudi et al. 86] Baroudi, J.; Olson, M.; & Ives, B. “An Empirical Study of the Impact of User Involvement on System Usage and Information Satisfaction.” *Communications of the ACM 29,* 3 (March 1986): 232-238.

[Bate et al. 95] Bate, R.; Kuhn, D.; Wells, C.; Armitage, J.; Clark, G.; Cusick, K.; Garcia, S.; Hanna, M.; Jones, R.; Malpass, P.; Minnich, I.; Pierson, H.; Powell, T.; & Reichner, A. *A Systems Engineering Capability Maturity Model, Version 1.1* (CMU/SEI-95-MM-003). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, November 1995. <http://www.sei.cmu.edu/publications/documents/95.reports/95.mm.003.html> (1995).

[Blalock 68] Blalock, H. M. “The Measurement Problem: A Gap between the Languages of Theory and Research,” 5-27. In *Methodology in Social Research.* Edited by H. M. Blalock and A. B. Blalock. New York: McGraw-Hill, 1968.

[Bollinger & McGowan 91] Bollinger, T. & McGowan, C. “A Critical Look at Software Capability Evaluation.” *IEEE Software 8,* 4 (July 1991): 25-41.

[Byrnes & Phillips 96] Byrnes, P. & Phillips, M. *Software Capability Evaluation: Method Description, Version 3.0* (CMU/SEI-96-TR-002, ADA309160). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, April 1996. <http://www.sei.cmu.edu/publications /documents/96.reports/96.tr.002.html> (1996).

[Carmines & Zeller 79] Carmines, E. G. & Zeller, R.A. *Reliability and Validity Assessment.* Sage University Paper Series on Quantitative Applications in Social Sciences. Beverly Hills, CA: Sage Publications, 1979.

[Cattell 66] Cattell, R. B. "The Scree Test for the Number of Factors." *Multivariate Behavioral Research 1*, 2 (April 1966): 245-276.

[Clark 97] Clark, B. *The Effect of Software Process Maturity on Software Development Effort*. PhD Thesis, University of Southern California, 1997.

[Cliff 88] Cliff, N. R. "The Eigenvalues-Greater-Than-One Rule and Reliability of Components." *Psychological Bulletin 103*, 2 (March 1988): 276-279.

[Coffman & Thompson 97] Coffman, A. & Thompson, K. "Air Force Software Process Improvement Report." *Crosstalk 10*, 1 (January 1997): 25-27.

[Comrey 73] Comrey, A. *A First Course on Factor Analysis*. London: Academic Press, 1973.

[Cooper & Fisher 02] Cooper, J. & Fisher, M. *Software Acquisition Capability Maturity Model (SA-CMM), Version 1.03* (CMU/SEI-2002-TR-010, ADA39794). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, March 2002. <http://www.sei.cmu.edu/publications/documents/02.reports /02tr010.html> (2002).

[Cronbach 51] Cronbach, L. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika 16,* 3 (September 1951): 297-334.

[Curtis 96] Curtis, B. “Factor Structure of the CMM and Other Latent Issues.” *Proceedings of the 1996 SEPG Conference*. Atlantic City, New Jersey, May 20-23, 1996. Pittsburgh, PA:, Carnegie Mellon University, Software Engineering Institute, 1996.

[Curtis et al. 01] Curtis, B.; Hefley, B.; & Miller, S. *People Capability Maturity Model (P-CMM), Version 2* (CMU/SEI-2001-MM-01). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, November 1995. <http://www.sei.cmu.edu/publications/documents/01.reports/01mm001.html> (2001).

[Davis 89] Davis, F. “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology.” *Management Information Systems Quarterly 13*, 3 (September 1989): 319-340.

[Dunaway 96] Dunaway, D. *CMM-Based Appraisal for Internal Process Improvement (CBA IPI) Lead Assessor’s Guide, V1.1* (CMU/SEI-96-HB-003). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1996.

[Dunaway & Masters 96] Dunaway, D. & Masters, S. *CMM-Based Appraisal for Internal Process Improvement (CBA IPI): Method Description* (CMU/SEI-96-TR-007, ADA307934). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, April 1996. <http://www.sei.cmu.edu/publications/documents/96.reports/96-tr-007/96-tr-007-abstract.html> (1996).

[Dunaway & Masters 01] Dunaway, D. & Masters, S. *CMM-Based Appraisal for Internal Process Improvement (CBA IPI): Method Description* (CMU/SEI-2001-TR-033). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, Nov. 2001. <http://www.sei.cmu.edu/publications/documents/01.reports/01tr033.html> (2001).

[Dunaway & Baker 01] Dunaway, D. & Baker, M. *Analysis of CMM-Based Appraisal for Internal Process Improvement (CBA IPI) Assessment Feedback* (CMU/SEI-2001-TR-021). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, November 2001. <http://www.sei.cmu.edu/publications/documents/01.reports/01tr021.html> (2001).

[Dybå 00] Dybå, T. "An Instrument for Measuring the Key Factors of Success in Software Process Improvement." *Empirical Software Engineering: An International Journal 5*, 4 (December 2000): 357-390.

[Efron & Tibshirani 93] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. New York, N.Y.: Chapman and Hall, 1993.

[El-Emam & Goldenson 95] El-Emam, K. & Goldenson, D. "SPICE: An Empiricist's Perspective," 84-97. *Proceedings of the Second IEEE International Software Engineering Standards Symposium*. Montreal, Quebec, Canada, August 21-25, 1995. Los Alamitos, CA: IEEE Computer Society Press, 1995.

[El-Emam & Madhavji 95] El-Emam, K. & Madhavji, N. "The Reliability of Measuring Organizational Maturity. *Software Process Improvement and Practice 1*, 1 (September 1995): 3-25.

[El-Emam et al. 97] El-Emam, K.; Smith, B.; & Fusaro, P. "Modeling the Reliability of SPICE Based Assessments," 69-82. *Proceedings of the 3rd IEEE International Software Engineering Standards Symposium*. Walnut Creek, California, June 1-6, 1997. Los Alamitos, CA: IEEE Computer Society Press, 1997.

[El-Emam 98] El-Emam, K. "The Internal Consistency of the ISO/IEC 15504 Software Process Capability Scale," 72-81. *Proceedings of the 5th International Symposium on Software Metrics*. Bethesda, Maryland, November 20-21, 1997. Los Alamitos, CA: IEEE Computer Society, 1998.

[El-Emam et al. 98] El-Emam, K.; Simon, J.-M; Rousseau, S.; & Jacquet, E. "Cost Implications of Interrater Agreement for Software Process Assessment," 38-51. *Proceedings of the 5th International Symposium on Software Metrics.* Bethesda, Maryland, November 20-21, 1997. Los Alamitos, CA: IEEE Computer Society, 1998.

[El-Emam 99] El-Emam, K. "Benchmarking Kappa: Interrater Agreement in Software Process Assessments." *Empirical Software Engineering: An International Journal 4*, 2: (June 1999): 113-133.

[El-Emam & Garro 00] El-Emam, K. & Garro, I. Estimating the Extent of Standards Use: The Case of ISO/IEC 15504." *Journal of Systems and Software 53*, 2 (August 2000): 137-143.

[El-Emam & Goldenson 00] El-Emam, K. & Goldenson, D. "An Empirical Review of Software Process Assessments." *Advances in Computers 53* (2000): 319-423.

[El-Emam & Birk 00a] El-Emam, K. & Birk, A. "Validating the ISO/IEC 15504 Measure of Software Development Process Capability." *Journal of Systems and Software 51*, 2 (April 2000): 119-149.

[El-Emam & Birk 00b] El-Emam, K. & Birk, A. "Validating the ISO/IEC 15504 Measure of Software Requirement Analysis Process Capability." *IEEE Transactions on Software Engineering 26*, 6 (June 2000): 541-566.

[El-Emam et al. 01] El-Emam, K.; Goldenson, D.; McCurley, J.; & Herbsleb, J. "Modeling the Likelihood of Software Process Improvement: An Exploratory Study." *Empirical Software Engineering: An International Journal 6*, 3 (September 2001): 207-229.

[Eum et al. 01] Eum, S.-H.; Cho, S.-C.; & Jung, H.-W. "Factors Affecting User Satisfaction with High Speed Internet Service." *Proceedings of the 14th International Conference on Software & Systems Engineering & their Applications* (CD-ROM). Paris, France, December 4-6, 2001.

[Fayad & Laitinen 97] Fayad, M. & Laitinen, M. "Process Assessment Considered Wasteful." *Communications of the ACM 40*, 11 (November 1997): 125-128.

[Fusaro et al. 98] Fusaro, P.; El-Emam, K.; & Smith, B. "The Internal Consistencies of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension." *Empirical Software Engineering: An International Journal 3*, 2 (January 1998): 179-201.

[Galletta & Lederer 89] Galletta, D. & Lederer, A. "Some Cautions on the Measurement of User Information Satisfaction." *Decision Sciences 20*, 3 (Summer 1989): 419-438.

[Goodhue 98] Goodhue, D. "Development and Measurement Validity of a Task-Technology Fit Instrument for User Evaluation of Information Systems." *Decision Sciences 29*, 1 (Winter 1998): 105-138.

[Gray & Smith 98] Gray, E. & Smith, W. "On the Limitations of Software Process Assessment and the Recognition of a Required Re-Orientation for Global Process Improvement." *Software Quality Journal 7*, 1 (March 1998): 21-34.

[Herbsleb et al. 97] Herbsleb, J.; Zubrow, D.; Goldenson, D.; Hayes, W.; & Paulk, M. "Software Quality and the Capability Maturity Model." *Communications of the ACM 40*, 6 (June 1997): 30-40.

[Humphrey 87] Humphrey, W. *Characterizing the Software Process: A Maturity Framework* (CMU/SEI-87-TR-11, ADA182895). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, June 1987.

[Humphrey & Sweet 87] Humphrey, W. & Sweet, W. *A Method for Assessing the Software Engineering Capability of Contractors* (CMU/SEI-87-TR-23, ADA187320). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, September 1987.

[Humphrey 88] Humphrey, W. "Characterizing the Software Process." *IEEE Software 5*, 2 (March 1988): 73-79.

[Humphrey 89] Humphrey, W. *Managing the Software Process*, Reading, MA: Addison-Wesley, 1989.

[ISO 98] ISO/IEC TR2 15504. *Information Technology–Software Process Assessment: Part 1–Part 9* (Part 5 was published in 1999). Geneva, Switzerland: ISO, 1998.

[ISO 99] ISO/IEC JTC1 Directive. *Procedures for the Technical Work of ISO/IEC JTC1*. Geneva, Switzerland. <http://www.jtc1.org /directives/toc.htm> (1999).

[ISO 01] ISO/IEC FCD 15939. *Information Technology–Software Engineering - Software Measurement.* Geneva, Switzerland: ISO, 2001.

[ISO/WG10 98] *Phase 2 Trials Interim Report*, Version 0.08. ISO/IEC JTC1/SC7/WG10. February 1998.

[ISO/WG10 99] *SPICE Phase 2 Trials Final Report*, Vol. 1. ISO/IEC JTC1/SC7/WG10. February 1999.

[ISO/WG10 01] *SPICE Network*. ISO/IEC JTC1/SC7/WG10. January 2001.

[Ives et al. 83] Ives, B.; Olson, M.; & Baroudi, J. "The Measurement of User Information Satisfaction." *Communications of the ACM 26*, 10 (October 1983): 785-793.

[Jiang et al. 00] Jiang, J.; Klein, G.; & Crampton, S. "A Note on SERVQUAL Reliability and Validity in Information System Service Quality Management." *Decision Sciences 31*, 3 (Summer 2000): 725-744.

[Jung & Hunter 01] Jung, H.-W. & Hunter, R. "The Relationship between ISO/IEC 15504 Process Capability Levels, ISO 9001 Certification and Organization Size: An Empirical Study." *Journal of Systems and Software 59*, 1 (October 2001): 23-41.

[Jung et al. 01] Jung, H.-W.; Hunter, R.; Goldenson D.; & El-Emam K. "Findings from Phase 2 of the SPICE Trials." *Software Process Improvement and Practice 6*, 2 (December 2001): 205-242.

[Jung & Hunter 02] Jung, H.-W. & Hunter, R. "An Evaluation of the SPICE Rating Scale with Regard to the Internal Consistency of Capability Measurement," 105-115. *Proceedings of the SPICE 2002, The Second International SPICE Conference*, Venice, Italy, March 14-15, 2002. Rome, Italy: Nuovo Studio Tecna, 2002.

[Kaiser 70] Kaiser, H. F. "A Second Generation Little Jiffy." *Psychometrika 35*, 4 (December 1970): 401-415.

[Krishnan & Kellner 99] Krishnan, M. S. & Kellner, M. I. "Measuring Process Consistency: Implications for Reducing Software Defects." *IEEE Trans. on Software Engineering 25*, 6 (November 1999): 800-815.

[Kuder & Richardson 37] Kuder, G. F. & Richardson, M. W. "The Theory of the Estimation of Test Reliability," *Psychometrika 2*, 3 (September 1937): 151-160.

[Kuvaja 99] Kuvaja, P. "Bootstrap 3.0–A SPICE Conformant Software Process Assessment Methodology." *Software Quality Journal 8*, 1 (September 1999): 7-19.

[Lissitz & Green 75] Lissitz, R. W. & Green, S. B. "Effects of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology 60*, 1 (February 1975): 10-13.

[Lunneborg 00] Lunneborg, C. *Data Analysis by Resampling: Concepts and Applications*. Pacific Grove, CA: Duxbury, Thomson Learning, 2000.

[Mathsoft 99] Mathsoft, Inc. *S-Plus 2000 User Guide*. Seattle, WA: Mathsoft, Inc., Data Analysis Products Division, 1999.

[Masters & Bothwell 95] Masters, S. & Bothwell, C. *CMM Appraisal Framework, Version 1.0* (CMU/SEI-95-TR-001, ADA 293300). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1995. <http://www.sei.cmu.edu/publications/documents/95.reports/95-tr-001/95-tr-001-abstract.html> (1995).

[McIver & Carmines 81] McIver, J. P. & Carmines, E. G. *Unidimensional Scaling*. Sage University Paper Series on Quantitative Applications in Social Sciences. Newbury Park, CA: Sage Publications, 1981.

[Mirani & Lederer 98] Mirani, R. & Lederer, A. "An Instrument for Accessing the Organizational Benefits of IS Projects." *Decision Sciences 29*, 4 (Fall 1998): 803-838.

[Montgomery et al. 98] Montgomery, D. C.; Runger, G. C.; & Hubele, N. F. *Engineering Statistics*. New York, N.Y: John Wiley & Sons, Inc., 1998.

[Nunnally & Bernstein 94] Nunnally, J. C. & Bernstein, I. H. *Psychometric Theory*. New York, NY: McGraw-Hill, Inc., 1994.

[Olson et al. 89] Olson, T.; Humphrey, W.; & Kitson, D. *Conducting SEI-Assisted Software Process Assessments* (CMU/SEI-89-TR-007, ADA219065). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 1989. <http://www.sei.cmu.edu/publications/documents/89.reports/89.tr.007.html> (1989).

[Paulk et al. 91] Paulk, M.; Curtis, B.; & Chrissis, M. B. *Capability Maturity Model for Software* (CMU/SEI-91-TR-24, ADA240603). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, August 1991.

[Paulk et al. 93a] Paulk, M.; Curtis, B; Chrissis, M. B.; & Weber, C. *Capability Maturity Model for Software, Version 1.1* (CMU/SEI-93-TR-024, ADA 263403). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, February 1993. <http://www.sei.cmu.edu /publications/documents/93.reports/93.tr.024.html> (1993).

[Paulk et al. 93b] Paulk, M.; Weber, C.; Garcia, S.; Chrissis, M. B.; & Bush, M. *Key Practices of the Capability Maturity Model, Version 1.1* (CMU/SEI-93-TR-025, ADA263432). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, February 1993. <http://www.sei.cmu.edu/publications/documents/93.reports /93.tr.025.html> (1993).

[Paulk et al. 93c] Paulk, M.; Curtis, B.; Chrissis, M. B.; & Weber, C. "Capability Maturity Model, Version 1.1." *IEEE Software 10*, 4 (July 1993): 18-27.

[Paulk et al. 94] Paulk, M.; Webber, C.; Curtis, B.; & Chrissis, M. B. *The Capability Maturity Model: Guidelines for Improving the Software Process.* New York, NY: Addison-Wesley, 1994.

[Paulk 95] Paulk, M. "The Evolution of the SEI's Capability Maturity Model for Software." *Software Process Improvement and Practice, Pilot Issue 1*, PI (August 1995): 3-15.

[Paulk 99] Paulk, M. "Analyzing the Conceptual Relationship Between ISO/IEC 15504 (Software Process Assessment) and the Capability Maturity Model for Software," 293-303. Proceedings, Ninth International Conference on Software Quality, October 4-6, 1999, Cambridge, Massachusetts. Milwaukee, WI: American Society for Quality Control, 1999.

[Rout et al. 98] Route, T.; Tuffley, A.; & Hodgen, B. "Process Capability in the Australian Software Industry: Results from the SPICE Trials," 12-19. *Australian Software Engineering Conference (ASWEC'98*), Adelaide, Australia, 1998. Los Alamitos, CA: IEEE Computer Society, 1998.

[Saiedian & Kuzara 95] Saiedian, H. & Kuzara, R. "SEI Capability Maturity Models' Impact on Contractors." *Computer 28*, 1 (January 1995): 16-26.

[SEI 01] *Standard CMMI Appraisal Method for Process Improvement (SCAMPI), Version 1.1: Method Definition Document* (CMU/SEI-2001-HB-001). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 2002. <http://www.sei.cmu.edu /publications/documents/01.reports/01hb001.html> (2001).

[SEI 02] *CMMI for Systems Engineering/Software Engineering/Integrated Product and Process Development/Supplier Sourcing, Version 1.1, Continuous Representation* (CMMI-SE/SW/IPPD/SS, V1.1) (Continuous: CMU/SEI-2002-TR-011) and *Staged Representation* (CMU/SEI-2002-TR-012). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, 2002. <http://www.sei.cmu.edu/cmmi/models/models.html> (2002).

[Sethi & King 91] Sethi, V. & King, W. "Construct Measurement in Information Systems Research: An Illustration in Strategic Systems." *Decision Sciences 22*, 3 (July 1991): 455-472.

[Sharma 96] Sharma, S. *Applied Multivariate Techniques*. New York, NY: John Wiley & Sons, Inc., 1996.

[Spector 92] Spector, P. F. *Summated Rating Scale Construction: An Introduction.* Sage University Paper Series on Quantitative Applications in Social Sciences. Newbury Park, CA: Sage Publications, 1992.

[Srinivasan 85] Srinivasan, A. "Alternative Measures of System Effectiveness: Associations and Implications." *Management Information Systems Quarterly 9*, 3 (September 1985): 243-253.

[Subramanian & Nilakanta 94] Subramanian, A. & Nilakanta, S. "Measurement: A Blueprint for Theory Building in MIS." *Information and Management 26*, 1 (January 1994): 13-20.

[Tait & Vessey 88] Tait, P. & Vessey, I. "The Effect of User Involvement on System Success: A Contingency Approach." *Management Information Systems Quarterly 12*, 1 (March 1988): 91-108.

[Torkzadeh & Doll 94] Torkzadeh, G. & Doll, W. "The Test-Retest Reliability of User Involvement Instrument." *Information and Management 26*, 1 (January 1994): 21-31.

[Trochim 01] Trochim, W. M. The Research Methods Knowledge Base, 2nd Edition. <http://trochim.human.cornell.edu/kb/index.htm> (2001).

[Weber et al. 91] Weber, C.; Paulk, M.; Wise, C.; & Withey, J. *Key Practices of the Capability Maturity Model* (CMU/SEI-91-TR-25, ADA240604). Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, August 1991.

[Van Dyke et al. 99] Van Dyke, T.; Prybutok, V.; & Kappelman, L. "Cautions on the SERVQUAL Measure to Access the Quality of Information Systems Services." *Decision Sciences 30*, 3 (Summer 1999): 877-891.

[Van De Ven & Ferry 80] Van De Ven, A. H. & Ferry, D. L. *Measuring and Assessing Organizations*. New York, NY: John Wiley & Sons, Inc., 1980.

[Zeller & Carmines 80] Zeller, R. A. & Carmines, E. G. *Measurement in the Social Sciences: The Link Between Theory and Data.* Cambridge, UK: Cambridge University Press, 1980.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. **AGENCY USE ONLY** (Leave Blank)	2. **REPORT DATE** December 2002	3. **REPORT TYPE AND DATES COVERED** Final
4. TITLE AND SUBTITLE The Internal Consistency of Key Process Areas in the Capability Maturity Model® (CMM®) for Software (SW-CMM)		5. FUNDING NUMBERS F19628-00-C-0003
6. **AUTHOR(S)** Ho-Won Jung, Dennis R. Goldenson		
7. **PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)** Software Engineering Institute Carnegie Mellon University Pittsburgh, PA 15213		8. **PERFORMING ORGANIZATION REPORT NUMBER** CMU/SEI-2002-TR-037
9. **SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)** HQ ESC/XPK 5 Eglin Street Hanscom AFB, MA 01731-2116		10. **SPONSORING/MONITORING AGENCY REPORT NUMBER**
11. **SUPPLEMENTARY NOTES**		
12A **DISTRIBUTION/AVAILABILITY STATEMENT** Unclassified/Unlimited, DTIC, NTIS		12B **DISTRIBUTION CODE**

13. **ABSTRACT (MAXIMUM 200 WORDS)**

Evaluating the reliability of maturity level ratings is crucial for providing confidence in the results of software process assessments. This report examines the dimensions underlying the maturity construct in the Capability Maturity Model® (CMM®) for Software (SW-CMM) and then estimates the internal consistency (reliability) of each dimension. The analysis is based on 676 CMM-Based Appraisal for Internal Process Improvement (CBA IPI) assessments conducted during the period of January 2000 through April 2002. The results suggest that the SW-CMM maturity is a three-dimensional construct, with “Project Implementation” representing the maturity level 2 key process areas (KPAs), “Organization Implementation” representing the maturity level 3 KPAs, and “Quantitative Process Management” representing the KPAs at both maturity levels 4 and 5. The internal consistency for each of the three dimensions as estimated by Cronbach’s alpha exceeds the recommended value of 0.9. Although more should be learned about the distinctions between maturity levels 4 and 5, the internal consistency of those KPAs is comparable to those at levels 2 and 3.

14. **SUBJECT TERMS** SW-CMM, CBA IPI, internal consistency, dimensionality, Cronbach’s alpha, factor analysis, bootstrap resampling			15. **NUMBER OF PAGES** 74
16. **PRICE CODE**			
17. **SECURITY CLASSIFICATION OF REPORT** Unclassified	18. **SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	19. **SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	20. **LIMITATION OF ABSTRACT** UL

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. Z39-18 298-102

® Capability Maturity Model and CMM are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.