National Research Council Canada
Institute for Information Technology
Conseil national de recherches Canada
Institut de Technologie de l'information
ERB-1072

NRC·CNRC

Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability

Khaled El Emam and Andreas Birk
February 1999

Canada
NRC 43619

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de Technologie de l'information

Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability

Khaled El Emam and Andreas Birk
February 1999

Validating the ISO/IEC 15504 Measure of Software Requirements Analysis Process Capability

Khaled El Emam[1]
National Research Council, Canada
Institute for Information Technology
Building M-50, Montreal Road
Ottawa, Ontario
Canada K1A OR6
Khaled.El-Emam@iit.nrc.ca

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

Abstract

ISO/IEC 15504 is an emerging international standard on software process assessment. It defines a number of software engineering processes, and a scale for measuring their capability. One of the defined processes is software requirements analysis (SRA). A basic premise of the measurement scale is that higher process capability is associated with better project performance (i.e., predictive validity). This paper describes an empirical study that evaluates the predictive validity of SRA process capability. Assessments using ISO/IEC 15504 were conducted on 56 projects world wide over a period of two years. Performance measures on each project were also collected using questionnaires, such as the ability to meet budget commitments and staff productivity. The results provide strong evidence of predictive validity for the SRA process capability measure used in ISO/IEC 15504, but only for organisations with more than 50 IT Staff. Specifically, a strong relationship was found between the implementation of requirements analysis practices as defined in ISO/IEC 15504 and the productivity of software projects. For smaller organisations evidence of predictive validity was rather weak. This can be interpreted in a number of different ways: that the measure of capability is not suitable for small organisations, or that the SRA process capability has less affect on project performance for small organisations.

1 Introduction

Assessments of software projects indicate that 80% of MIS projects are at risk of creeping user requirements; and so are 70% of military projects, and 45% of contract or outsourced projects [44]. Moreover, a recent survey of European software organisations identified that more than 40% perceived that they had major problems in managing customer requirements, and more than 50% perceived that they had major problems in the area of requirements specification [42]. In addition, these were the two areas with the greatest perceived problems out of all the surveyed areas[2]. Another survey, also conducted in Europe, identified adoption levels of requirements engineering practices that are consistently smaller than 60%, such as procedures for ensuring appropriate levels of user/customer/marketing input (59% adoption), procedures for controlling changes to requirements, designs and documentation (58% adoption), tools for requirements traceability (22% adoption), and prototyping for validating requirements (57% adoption) [19]. Given this state of affairs, it would seem that further effort is necessary to improve requirements engineering practices.

A commonly used paradigm for improving software engineering practices in general is the benchmarking paradigm [9]. This involves identifying an 'excellent' organisation or project and documenting its

[1] Work done by El Emam was partially completed while he was at the Fraunhofer Institute for Experimental Software Engineering, Kaiserslautern, Germany.

[2] The other areas were "Documentation", "Software and System Test", "Lack of Quality System", "Project Management", "Lack of Standards", "System Analysis and Design", "Configuration Management", "Software Installation and Support", and "Program Coding".

practices. It is then assumed that if a less-proficient organisation or project adopts the practices of the excellent one, it will also become excellent. Such best practices are commonly codified in an assessment model, such as the SW-CMM[3] [81] or the emerging ISO/IEC 15504 international standard [24]. These assessment models also order the practices in a recommended sequence of implementation, hence providing a predefined improvement path.[4]

Some of these models include the requirements engineering process within their scope. Hence they define what are believed to be best requirements engineering practices. For instance, the SW-CMM has a Key Process Area (KPA) on requirements management [81]. The emerging ISO/IEC 15504 international standard defines a "develop software requirements" process [24].[5] In addition, a maturity model specific to requirements engineering has been defined which includes an extensive catalogue of requirements engineering practices organised in a recommended order of implementation [85].

Improvement following the benchmarking paradigm almost always involves a software process assessment (SPA).[6] A SPA provides a quantitative score reflecting the extent of an organisation's or project's implementation of the best practices defined in the assessment model. The more of these best practices that are adopted, the higher this score is expected to be. The obtained score provides a baseline of current implementation of best practices, serves as a basis for making process improvement investment decisions, and also provides a means of tracking improvement efforts.[7]

A basic premise of this approach is that the quantitative score from the assessment is associated with the performance of the organisation or project. Therefore, improving the requirements engineering practices according to an assessment model is expected to subsequently improve the performance. This is termed the *predictive validity* of the process capability score. Empirically validating the verisimilitude of such a premise is of practical importance since substantial process improvement investments are made by organisations guided by the assessment models.

While there have been some correlational studies that substantiate the above premise, these tended to evaluate composite process capability scores across multiple different processes, but have not provided results that are specific to the requirements engineering process. Therefore, thus far the relationship between the assessment score and performance remains a premise that enjoys weak empirical support for requirements engineering practices. The implication then is that it is not possible to substantiate claims that improvement by adopting requirements engineering practices stipulated in the assessment models really results in performance improvements.

In this paper we empirically investigate the relationship between the capability of the software requirements analysis (SRA) process as defined in the emerging ISO/IEC 15504 international standard and the performance of software projects. The study was conducted in the context of the SPICE Trials, which is an international effort to empirically evaluate the emerging international standard worldwide. To our knowledge, this is the first study to evaluate the predictive validity of SRA process capability using an internationally standardised measure of process capability.

Briefly, our results indicate that for large organisations, SRA process capability as measured in ISO/IEC 15504 is related to project productivity. This means that improvements in SRA process capability are associated with a reduction in the cost of software projects. This is interpreted to be due to a reduction in rework during the project. However, no relationship was found with other measures of performance, nor

---

[3] The Capability Maturity Model for Software.

[4] The logic of this sequencing is that this is the natural evolutionary order in which, historically, software organisations improve [40], and that practices early in the sequence are prerequisite foundations to ensure the stability and optimality of practices implemented later in the sequence [81].

[5] In this paper we only refer to the PDTR version of the ISO/IEC 15504 document set since this was the one used during our empirical study. The PDTR version reflects one of the stages that a document has to go through on the path to international standardisation. The PDTR version is described in detail in [24].

[6] Here we use the term "SPA" in the general sense, not in the sense of the SEI specific assessment method (which was also called a SPA).

[7] A recent survey of sponsors of assessments indicated that baselining process capability and tracking process improvement progress are two important reasons for conducting a SPA [27].

was there any relationship between SRA process capability and any of the peformance measures that were used for small organisations.

In the next section we provide the background to our study. This is followed in Section 3 with an overview of the ISO/IEC 15504 architecture and rating scheme that was used during our study. Section 4 details our research method, and Section 5 contains the results. We conclude the paper in Section 6 with a discussion of our results and directions for future research.

2 Background

A SPA can be considered as a measurement procedure. As with any measurement procedure, its validity must be demonstrated before one has confidence in its use. The validity of measurement is defined as the extent to which a measurement procedure is measuring what it is purporting to measure [47]. During the process of validating a measurement procedure one attempts to collect evidence to support the types of inferences that are to be drawn from measurement scores.
A basic premise of SPAs is that the resultant quantitative scores are associated with the performance of the project and/or organisation that is assessed. This premise consists of two parts:[8]

- that the practices defined in the assessment model are indeed good practices and their implementation will therefore result in improved performance
- that the quantitative assessment score is a true reflection of the extent to which these practices are implemented in the organisation or project; and therefore projects or organisations with higher assessment scores are likely to perform better.

Testing this premise can be considered as an evaluation of the *predictive validity* of the assessment measurement procedure [21].

In this section we review existing theoretical and empirical work on the measurement of SRA process capability and the predictive validity of such measures. However, first we present our terminology, and discuss some methodological issues in the evaluation of predictive validity.

2.1 Terminology

In this subsection we define the terminology that is used throughout the paper. This is to avoid confusion since the literature discusses predictive validity for different units of analysis.

[8] The fact that the premise behind the use of quantitative scores from SPAs consists of two parts means that if no empirical evidence is found to support the basic premise, then we would not know which part is at fault. For example, if we find that there is no relationship between the assessment score and performance it may be because:

- the practices are really not good practices, but the measurement procedure is accurately measuring their implementation, or
- the practices are really good practices, but the measurement procedure is not accurately measuring their implementation.

From a practical standpoint it does not matter which of the above two conclusions one draws since the practices and measurement procedure are always packaged and used together.

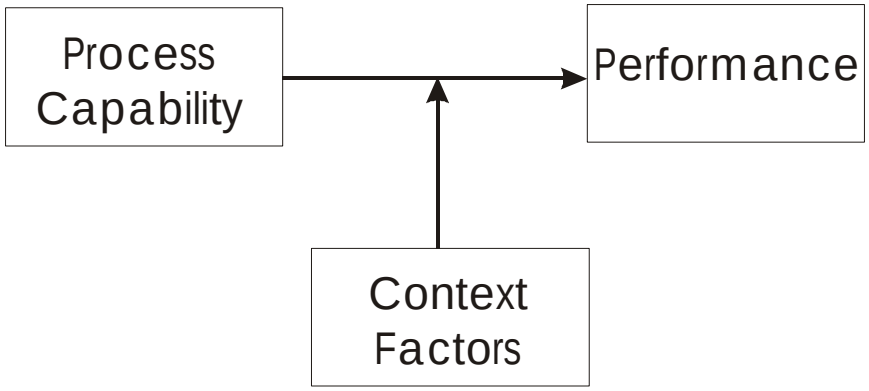

Figure 1: Theoretical model being tested in a predictive validity study of process capability.

A predictive validity study typically tests the hypothesised model shown in Figure 1. This shows that there is a relationship between process capability and performance, and that this relationship is dependent upon some context factors (i.e., the relationship functional form or direction may be different for different contexts, or may exist only for some contexts).

The hypothesised model can be tested for different units of analysis [35]. The three units of analysis are the life cycle process (e.g., the SRA process)[9], the project (which could be a composite of the capability of multiple life cycle processes of a single project, such as SRA and design), or the organisation (which could be a composite of the capability of the same or multiple processes across different projects). All of the three variables in the model can be measured at any one of these units of analysis. Therefore, in our review of the empirical literature we will precede each of the variables with its unit of analysis. For example, we will refer to *SRA* process capability, *project* process capability, *SRA* performance, and *project* performance.[10]

2.2 Theoretical Basis for Validating SRA Process Capability

Four existing models hypothesise benefits as requirements engineering process capability is improved. These are reviewed below.

Requirements Management is a KPA defined at Level 2 of the SW-CMM [81]. The goals of this KPA are:

- System requirements allocated to software are controlled to establish a baseline for software engineering and management use.
- Software plans, products and activities are kept consistent with the system requirements allocated to software.

Furthermore, the Software Product Engineering KPA defines software requirements analysis activities (namely activity 2) [83]. This KPA is defined at Level 3.

As organisations increase their organisational process capability, it is hypothesised that three types of benefits will accrue [63]:

- the differences between targeted results and actual results will decrease across projects,
- the variability of actual results around targeted results decreases, and
- costs decrease, development time shortens, and productivity and quality increase.

[9] One can make the distinction between organisational processes (e.g., the ISO/IEC 15504 “Engineer the Business” or “Provide Software Engineering Infrastructure” processes [24]) and project specific processes, as was done in a recent study [25]. In the current study we only focus on the SRA process, which is a project specific processes.

[10] A distinction has been made between measuring process capability, as in ISO/IEC 15504, and measuring organisational maturity, as in the SW-CMM [64]. According to our terminology, the SW-CMM would be measuring organisational process capability, although the measurement scheme may be quite different from the scheme used by another model, such as ISO/IEC 15504.

However, these benefits are not posited only for the requirements management KPA, but rather as a consequence of implementing combinations of practices.

The emerging ISO/IEC 15504 international standard, on the other hand, defines a set of processes, and a scale that can be used to evaluate the capability of each process separately [24] (details of the ISO/IEC 15504 architecture are provided in Section 3). The initial requirements for ISO/IEC 15504 state that an organisation's assessment results should reflect its ability to achieve productivity and/or development cycle time goals [24]. It is not clear however, whether this is hypothesised for each individual process, or for combinations of processes.

Somerville and Sawyer have defined a process capability model that is specialised for requirements engineering [85]. It comprises three levels: initial, repeatable, and defined. For each level, a set of practices and guidelines are provided which are more detailed than the corresponding elements of the SW-CMM and ISO/IEC 15504. The purpose of the capability model is to guide the implementation and improvement of advanced SRA practices in industry. Somerville and Sawyer argue that high process capability is very likely to result in better quality of SRA results. Although they do not explicitly relate SRA process capability to the performance of the overall organisation, they do specify the expected benefits of implementing each of the practices.

The Software Engineering Institute has published a so-called *Technology Reference Guide* [82], which is a collection and classification of software technologies. Its purpose is to foster technology dissemination and transfer. Each technology is classified according to processes in which it can be applied (*application taxonomy*) and according to qualities of software systems that can be expected as a result of applying the technology (*quality measures taxonomy*). The classifications have passed a comprehensive review by a large number of software engineering experts. This accumulated expert opinion can be used as another theoretical source on the impact of SRA processes on overall process performance. The technologies listed for the process categories *requirements engineering* and *requirements tracing* support the quality measures effectiveness and correctness, maintainability and understandability, as well as reusability. No particular effects are stated for system performance and organisational measures (e.g., cost of ownership for the developed system). If we include *cost estimation* within the scope of software requirements analysis, then there are a number of technologies listed that have an impact on productivity. As a conclusion from these classifications it can be expected that certain SRA practices have particular effects on certain performance variables. Since process capability is defined through the implementation of practices, some correlation between process capability and process performance can reasonably be expected.

Therefore, the existing literature does strongly suggest that there is a relationship between SRA process capability and performance. However, the models differ in their definitions of "good" SRA practices, in the expected benefits that they contend will accrue from their implementation, and also the former three in their process capability measurement schemes.

2.3 Issues in the Evaluation of Predictive Validity

Many previous empirical predictive validity studies used a composite measure of process capability. This means that they measured the capability of individual processes, and then these individual measures were aggregated to produce an overall project or organisational measure. To pre-empt ourselves, we show below that predictive validity studies that use composite scores can be of most value when the quantitative assessment results are used for supplier selection, but have weak utility in a process improvement context. Since our main focus is on improving SRA practices, this conclusion highlights the need for predictive validity studies that do not use composite measures, even if they include the SRA process as one of their components.

A common coefficient for the evaluation of predictive validity in general is the correlation coefficient [62]. It has also been used in the context of evaluating the predictive validity of project and organisational process capability measures [59][22]. We therefore use this coefficient in our study.

Let us assume that k different process capability measures for k different processes have been obtained, $x_1, x_2, x_3, \ldots, x_k$. We then construct a linear composite as follows:

$X = x_1 + x_2 + x_3 + \ldots + x_k$. Therefore, X is measuring overall process capability across different processes. Further, let us assume that the criterion (performance) variable is denoted by Y (e.g., this could be productivity), and that we have collected data from n entities (these may be projects or organisations, for example). Below we define some relationships in terms of population parameters. However, these can also be substituted with sample estimates.

Initially, we define some general relationships. The correlation between any two variables, a and b can be expressed as:

$$r_{ab} = \frac{s_{ab}}{s_a s_b} \qquad \textbf{Eqn. 1}$$

where s_{ab} is the covariance, and $s_a s_b$ is the product of the standard deviation of the a and b variables respectively.

Also, note that the mean of a composite variable X can be defined as:

$$\bar{X} = \frac{\sum (x_1 + \cdots + x_k)}{n} = \frac{\sum x_1}{n} + \cdots + \frac{\sum x_k}{n} = \bar{x}_1 + \cdots + \bar{x}_k \qquad \textbf{Eqn. 2}$$

We also need to define the variance of the X variable, which is (using Eqn. 2):

$$s_X^2 = \sum \frac{(X - \bar{X})^2}{n} = \sum \frac{\left((x_1 - \bar{x}_1) + (x_2 - \bar{x}_2) + \cdots + (x_k - \bar{x}_k)\right)^2}{n}$$

$$= \sum \frac{\left((x_1 - \bar{x}_1)^2 + \cdots + (x_k - \bar{x}_k)^2 + 2(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) + \cdots + 2(x_{k-1} - \bar{x}_{k-1})(x_k - \bar{x}_k)\right)}{n}$$

$$s_X^2 = s_1^2 + \cdots + s_k^2 + 2 s_1 s_2 r_{12} + \cdots + 2 s_{k-1} s_k r_{(k-1)k} \qquad \textbf{Eqn. 3}$$
$$= k \bar{s}_x^2 + k(k-1) \overline{s_x s_{x'} r_{xx'}}$$

where s_i^2 is the variance of component i of the X composite, and:

$$\bar{s}_x^2 = \frac{s_1^2 + \cdots + s_k^2}{k} \qquad \textbf{Eqn. 4}$$

which is the mean of the variances of the components of the X composite, and:

$$\overline{s_x s_{x'} r_{xx'}} = \frac{s_1 s_2 r_{12} + \cdots + s_{k-1} s_k r_{(k-1)k}}{\binom{k(k-1)}{2}} \qquad \textbf{Eqn. 5}$$

which is the mean of the $k(k-1)/2$ covariance terms in Eqn. 3, and $r_{xx'}$ is used to denote the correlation between any pair x and x' in the X composite.

With these preliminaries, now we can express the relationship between the composite and the criterion in terms of the variances and covariances of the components of the X composite.

The correlation between the composite process capability measure and the criterion (i.e., the predictive validity coefficient), r_{XY}, can be defined by:

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} \qquad \textbf{Eqn. 6}$$

We then have:

$$r_{XY} = \frac{\sum\left(\left(Y - \bar{Y}\right)\left((x_1 - \bar{x}_1) + \cdots + (x_k - \bar{x}_k)\right)\right)}{n s_Y s_X} \qquad \textbf{Eqn. 7}$$

$$= \frac{s_{Y1} + \cdots + s_{Yk}}{s_Y s_X} = \frac{s_Y s_1 r_{Y1} + \cdots + s_Y s_k r_{Yk}}{s_Y s_X} = \frac{s_1 r_{Y1} + \cdots + s_k r_{Yk}}{s_X}$$

If we convert all scores in the X composite into standard score form, then their variances become equal to one (i.e., $s_1 = \cdots = s_k = 1$), and using Eqn. 3, we can express Eqn. 7 as:

$$r_{XY} = \frac{r_{Y1} + \cdots + r_{Yk}}{\sqrt{k\bar{s}_x^2 + k(k-1)\overline{s_x s_{x'} r_{xx'}}}} = \frac{k\bar{r}_{Yx'}}{\sqrt{k + k(k-1)\bar{r}_{xx'}}} \qquad \textbf{Eqn. 8}$$

$$= \frac{\bar{r}_{Yx'}}{\sqrt{1/k + \left(\frac{(k-1)}{k}\right)\bar{r}_{xx'}}}$$

where $\bar{r}_{Yx'}$ is the average of the correlation coefficients between each of the components of the X composite and the criterion variable, and $\bar{r}_{xx'}$ is the average of the correlation coefficients between pairs of components making up the X composite.

Eqn. 8 tells us some interesting characteristics of the predictive validity coefficient between a process capability composite and some criterion variable. We explain these with reference to Figure 2:

- In general, the larger the number of components in the X composite the greater the predictive validity coefficient will be. Although this tends to plateau, there is still a slight increase. This means that the more processes that are included in a process capability composite measure, the greater the predictive validity, by definition[11].
- As can be seen in panel (a), as the average inter-component correlation ($\bar{r}_{xx'}$) decreases, the predictive validity coefficient increases. This means that by combining process capability scores that are measuring different processes that are not related to each other, one is effectively increasing predictive validity.
- As can be seen in panel (b), as the average correlation between each of the components and the criterion ($\bar{r}_{Yx'}$) increases, so does the predictive validity coefficient. This means that if only one of the process capability measures in a composite is strongly related to the criterion, and all of the others have a weak relationship, the average correlation could still be large enough to give an overall high predictive validity coefficient. Therefore, studies that demonstrate a nontrivial relationship between the composite and the criterion only tell us that there are some process capability measures (or even only one) in the composite that are related with the criterion. It by no means tells us that all of the measures of process capability in the composite are strongly related to the criterion. Furthermore, since it is a composite, we do not know which one of its components is strongly associated with the criterion.

[11] This assumes that additional process capability measures will maintain the same average inter-component correlation and the same average correlation with the criterion.

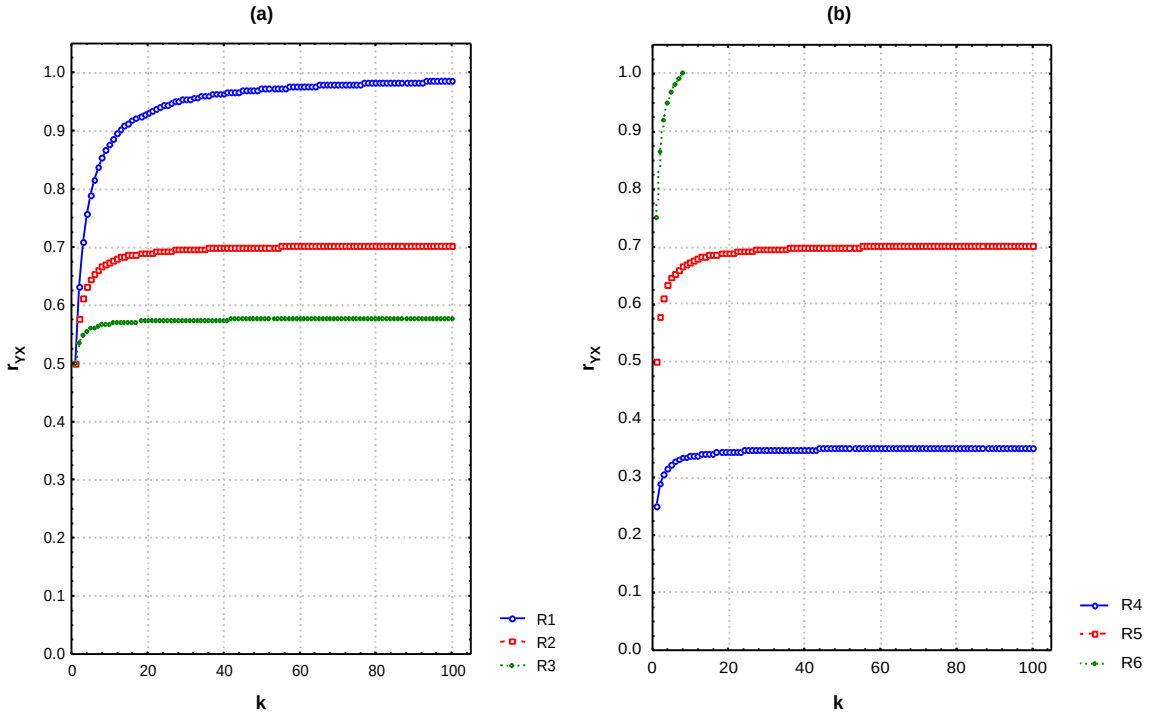

Figure 2: The two above panels show the behavior of the predictive validity coefficient as a function of the number of components in the X composite (k) as it varies up to 100. The plots in panel (a) are for $\bar{r}_{Yx'}$ fixed at 0.5, and $\bar{r}_{xx'}$ varies as follows: plot R1 is for $\bar{r}_{xx'}$=0.25, plot R2 is for $\bar{r}_{xx'}$=0.5, plot R3 is for $\bar{r}_{xx'}$=0.75. The plots in panel (b) are for $\bar{r}_{xx'}$ fixed at 0.5, and $\bar{r}_{Yx'}$ varies as follows: plot R4 is for $\bar{r}_{Yx'}$=0.25, plot R5 is for $\bar{r}_{Yx'}$=0.5, plot R6 is for $\bar{r}_{xx'}$=0.75.

We now define two contexts where process capability measures can be used: a capability determination context, and a software process improvement context.

In a capability determination context the assessment scores are used to select suppliers. In a crude implementation, one selects the supplier with the highest process capability. In a more refined implementation of the decision process (e.g., see [11]), one uses the process capability score(s) as only one input, albeit it may be weighted heavily.

The primary interest in a capability determination context is in a process capability measure that is monotonically related with performance. It does not really matter how this process capability measure is composed as long as it is a good predictor of performance. To achieve this, according to Eqn. 8, one should include many processes in a composite process capability measure, they ought to be as different as possible (i.e., with low inter-correlations), and each one is highly related with performance.

If nontrivial evidence of predictive validity is found, then this has achieved the purpose. We can use the composite as an indicator of performance, and select the supplier with the largest composite process capability. However, this, at best, only indicates that there is something in that composite that is associated with the criterion (not that the capability of all the processes in the composite are related to the criterion).[12] While this may be acceptable when selecting a supplier, it is not sufficient in a process

[12] To make the point more concrete, let us assume that we have 5 process capability measures in the X composite, and that the average inter-component correlation ($\bar{r}_{xx'}$) is 0.3. Further, let's say that a predictive validity study found the correlation between the

improvement context, where one also desires to be able to identify specific processes that ought to be improved. Given that substantial investments are made based on the stipulations of assessment models, such evidence is not actually very useful in a practical sense for process improvement.[13]

Since in the context of process improvement it is individual processes that are improved, it is important to demonstrate predictive validity for capability measures of individual processes. Consequently, demonstrating predictive validity for a composite measure does not necessarily provide evidence of predictive validity of its individual process capability measures as demonstrated above. This argument then provides a context for the empirical literature review that follows.

2.4 Evidence of the Predictive Validity of SRA Process Capability

To our knowledge, no empirical evidence exists supporting the predictive validity of SRA process capability measures as defined by Sommerville and Sawyer nor as defined in ISO/IEC 15504. The Technology Reference Guide is based largely on expert judgement. Nevertheless, there have been studies of predictive validity based on the SW-CMM and other models. The following review indicates that of the previous validation studies that incorporate the SRA process or SRA practices within their scope they utilised composite measures, and the remaining studies that considered individual process did not include the SRA process within their scope.

Two classes of empirical studies have been conducted and reported thus far: case studies and correlational studies [35]. Case studies describe the experiences of a single organisation (or a small number of selected organisations) and the benefits it gained from increasing its process capability. Case studies are most useful for showing that there are organisations that have benefited from increased process capability. Examples of these are reported in [41][38][17][18][92][2][57][8][54] (also see [49] for a recent review). However, in this context, case studies have a methodological disadvantage that makes it difficult to generalise the results from a single case study or even a small number of case studies. Case studies tend to suffer from a selection bias because:

- Organisations that have not shown any process improvement or have even regressed will be highly unlikely to publicise their results, so case studies tend to show mainly success stories (e.g., all the references to case studies above are success stories), and
- The majority of organisations do not collect objective process and product data (e.g., on defect levels, or even keep accurate effort records). Only organisations that have made improvements and reached a reasonable level of maturity will have the actual objective data to demonstrate improvements (in productivity, quality, or return on investment). Therefore failures and non-movers are less likely to be considered as viable case studies due to the lack of data.

With correlational studies, one collects data from a number of organisations or projects and investigates relationships between process capability and performance statistically. Correlational studies are useful for showing whether a general association exists between increased capability and performance, and under what conditions.

There have been a few correlational studies in the past that evaluated the predictive validity of various process capability measures. For example, Goldenson and Herbsleb [34] evaluated the relationship between SW-CMM capability scores and organisational performance measures. They surveyed

---

composite and the performance variable to be 0.6 (r_{XY}). What does this tell us ? The predictive validity coefficient r_{XY} is certainly of a respectable magnitude. Using Eqn. 8, we can calculate that the average correlation between each of the components of the composite and the criterion ($\overline{r}_{Yx'}$) is 0.39. Consider a situation where only one of the components of the composite has a correlation of 0.95 with the criterion, and the remaining four components have a correlation of 0.25 with the criterion. This gives a $\overline{r}_{Yx'}$ of 0.39! In such a case, only one process capability measure within the component had quite a large correlation with performance, and the others had a rather small correlation, but the overall correlation of the composite with the criterion was respectable and indicative of good predictive validity. Therefore, a high composite correlation is not an indicator that each of its components is equally nor highly correlated with the criterion. In fact, in this case only one of the components had a high correlation with the criterion.

[13] Consider telling a sponsor of an assessment that they should invest in improving 5 processes because we have evidence that at least one of them is useful, but not necessarily all of them, and that we do not know which one.

individuals whose organisations have been assessed against the SW-CMM. The authors evaluated the benefits of higher process capability using subjective measures of performance. Organisations with higher capability tend to perform better on the following dimensions (respondents chose either the "excellent" or "good" response categories when asked to characterise their organisation's performance on these dimensions): ability to meet schedule, product quality, staff productivity, customer satisfaction, and staff morale. The relationship with the ability to meet budget commitments was not found to be statistically significant.

A more recent study considered the relationship between the implementation of the SW-CMM KPA's and delivered defects (after correcting for size and personnel capability) [50]. They found evidence that increasing process capability is negatively associated with delivered defects. Another correlational study investigated the benefits of moving up the maturity levels of the SW-CMM [28][53]. They obtained data from historic U.S. Air Force contracts. Two measures were considered: (a) cost performance index which evaluates deviations in actual vs. planned project cost, and (b) schedule performance index which evaluates the extent to which schedule has been over/under-run. Generally, the results show that higher maturity projects approach on-target cost, and on-target schedule. McGarry et al. [59] investigated the relationship between assessment scores using an adaptation of the SW-CMM process capability measures and project performance for fifteen projects within a single organisation. They did not find strong evidence of predictive validity, although they were all in the expected direction. Clark [10] investigated the relationship between satisfaction of SW-CMM goals and software project effort, after correcting for other factors such as size and personnel experience. His results indicate that the more KPAs are implemented, the less effort is consumed on projects. Jones presents the results of an analysis on the benefits of moving up the 7-level maturity scale of Software Productivity Research (SPR) Inc.'s proprietary model [46][45]. This data were collected from SPR's clients. His results indicate that as organisations move from Level 0 to Level 6 on the model they witness (compound totals): 350% increase in productivity, 90% reduction in defects, 70% reduction in schedules.

One can argue that the above studies provide ample evidence as to the predictive validity of process capability measures, although one study did not demonstrate that (the McGarry et al. study [59]). They all considered the SRA process in the composite measure of process capability, or elements thereof. However, since these studies used composite measures of process capability, it is not possible to determine whether SRA process capability is related to performance, and if so which specific performance measure(s) the SRA process capability is associated with.

Deephouse et al. evaluated the relationship between individual processes and project performance (as opposed to a composite across multiple processes) [16]. As would be expected, they found that evidence of predictive validity depends on the particular performance measure that is considered. However, this study did not focus specifically on the SRA process. One study by El Emam and Madhavji [22] evaluated the relationship between four dimensions of organisational process capability and the success of the requirements engineering process. Evidence of predictive validity was found for only one dimension. However, the organisational process capability dimensions were not specific to the SRA process.

There have been empirical investigations of specific SRA practices. Most notably is the literature on the benefits of user participation during the early phases of the life cycle and its affects on project and organisational performance, for example [30][48][61]. Another study investigated the affects of user participation in the requirements engineering process on the performance of the requirements engineering process itself [23]. Such studies are useful for identifying practices that are potentially beneficial to implement, however they do not address the issue of the capability of the whole of the SRA process.

As can be seen from the above review, no evidence exists that demonstrates the relationship between the capability of the SRA process and the performance of software projects. This means that we cannot substantiate claims that improving the capability of the SRA process will lead to any improvement in project performance, and we cannot be specific about which performance measures will be affected. Hence, the rationale for the current study.

2.5 Moderating Effects

A recent review of the empirical literature on software process assessments noted that existing evidence suggests that the extent to which a project's or organisation's performance improves due to the implementation of good software engineering practices (i.e., increasing process capability) is dependent on the context [26]. This highlights the need to consider the project and/or organisational context in predictive validity studies.

In our current study we consider the size of the organization as a context factor. This is not claimed to be the only context factor that ought to be considered, but is only one of the important ones that has been mentioned repeatedly in the literature. In general, it has been noted that the overall evidence remains equivocal as to which context factors should be considered in predictive validity studies [26].

Previous studies provide inconsistent results about the effect of organisational size. For example, there have been some concerns that the implementation of some of the practices in the CMM, such as a separate Quality Assurance function and formal documentation of policies and procedures, would be too costly for small organizations [6]. Therefore, the implementation of certain processes or process management practices may not be as cost-effective for small organisations as for large ones. However, a moderated analysis of the relationship between capability and requirements engineering process success (using the data set originally used in [22]) [26] found that organisational size does not affect predictive validity. This result is consistent with that found in [34] for organisation size and [16] for project size, but is at odds with the findings from [6].

To further confuse the issue, an earlier investigation [55] studied the relationship between the extent to which software development processes are standardised and MIS success.[14] It was found that standardisation of life cycle processes was associated with MIS success in smaller organisations but not in large ones. This is in contrast to the findings cited above. Therefore, it is not clear if and how organisation size moderates the benefits of process and the implementation of process management practices.

We therefore explicitly consider organisational size as a factor in our study to identify if the predictive validity results are different for different sized organisations.

2.6 Measures of Project Performance

To maintain comparability with previous studies, we define project performance in a similar manner. In the Goldenson and Herbsleb study [34] performance was defined in terms of six variables: customer satisfaction, ability to meet budget commitments, ability to meet schedule commitments, product quality, staff productivity, and staff morale / job satisfaction. We use these six variables, except that product quality is changed to "ability to satisfy specified requirements". We therefore define project performance in terms of the six variables summarised in Table 1. Deephouse et al. [16] consider software quality (defined as match between system capabilities and user requirements, ease of use, and extent of rework), and meeting targets (defined as within budget and on schedule). One can argue that if "ease of use" is not in the requirements then it ought not be a performance criterion, therefore we can consider it as being a component of satisfying specified requirements. Extent of rework can also be considered as a component of productivity since one would expect productivity to decrease with an increase in rework. Therefore, these performance measures are congruent with our performance measures, and it is clear that they represent important performance criteria for software projects.

[14] Process standardisation is a recurring theme in process capability measures.

Definition	Variable Name
Ability to meet budget commitments	BUDGET
Ability to meet schedule commitments	SCHEDULE
Ability to achieve customer satisfaction	CUSTOMER
Ability to satisfy specified requirements	REQUIREMENTS
Staff productivity	PRODUCTIVITY
Staff morale / job satisfaction	MORALE

Table 1: The criterion variables that were studied. These were evaluated for every project. The question was worded as follows: "How would you judge the process performance on the following characteristics …". The response categories were: "Excellent", "Good", "Fair", "Poor", and "Don't Know".

3 Overview of the ISO/IEC PDTR 15504 Rating Scheme

3.1 The Architecture

The architecture of ISO/IEC 15504 is two-dimensional as shown in Figure 3. One dimension consists of the processes that are actually assessed (the Process dimension) that are grouped into five categories. The second dimension consists of the capability scale that is used to evaluate the process capability (the Capability dimension). The same capability scale is used across all processes. The software requirements analysis process is defined in the Engineering process category in the Process dimension.

During an assessment it is not necessary to assess all the process in the process dimension. Indeed, an organisation can scope an assessment to cover only the subset of processes that are relevant for its business objectives. Therefore, not all organisations that conduct an assessment based on ISO/IEC 15504 will cover the requirements analysis process.

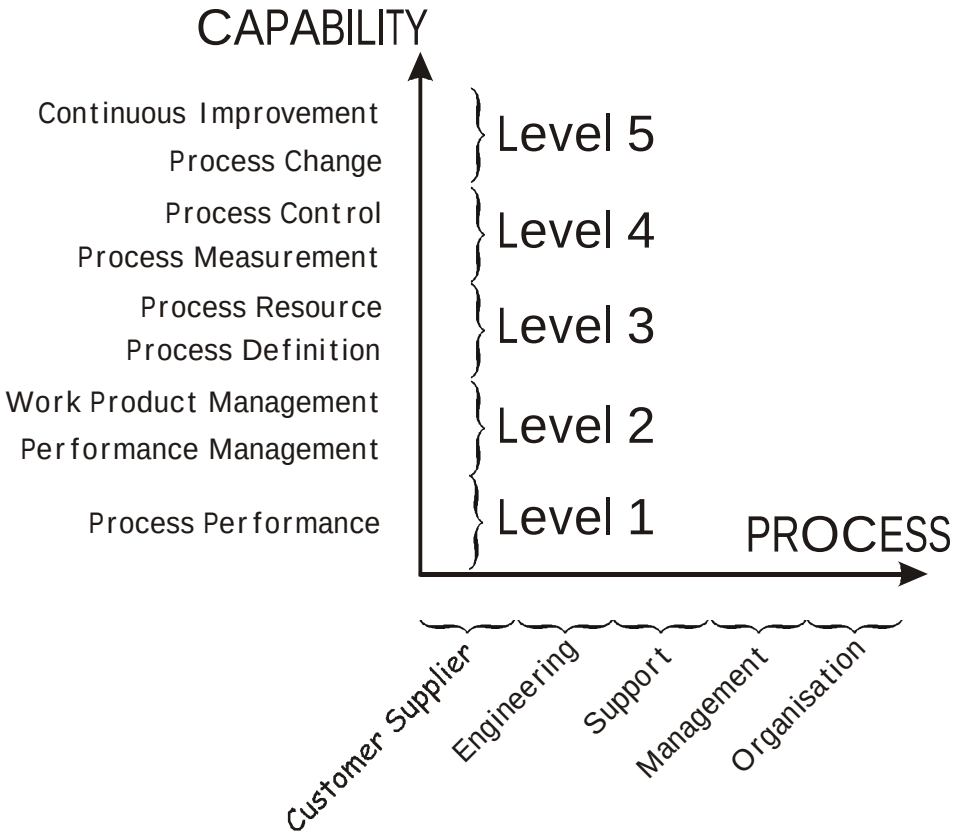

Figure 3: An overview of the ISO/IEC 15504 two dimensional architecture.

In ISO/IEC 15504, there are 5 levels of capability that can be rated, from Level 1 to Level 5. A Level 0 is also defined, but this is not rated directly. These 6 levels are shown in Table 2. In Level 1, one attribute is directly rated. There are 2 attributes in each of the remaining 4 levels. The attributes are also shown in Table 2 (also see [24]).

The rating scheme consists of a 4-point *achievement* scale for each attribute. The four points are designated as F, L, P, N for *Fully Achieved*, *Largely Achieved*, *Partially Achieved*, and *Not Achieved*. A summary of the definition for each of these response categories is given in Table 3.

The unit of rating in an ISO/IEC PDTR 15504 process assessment is the process instance. A process instance is defined as a singular instantiation of a process that is uniquely identifiable and about which information can be gathered in a repeatable manner [24].

ID	Title
Level 0	**Incomplete Process** There is general failure to attain the purpose of the process. There are no easily identifiable work products or outputs of the process.
Level 1	**Performed Process** The purpose of the process is generally achieved. The achievement may not be rigorously planned and tracked. Individuals within the organisation recognise that an action should be performed, and there is general agreement that this action is performed as and when required. There are identifiable work products for the process, and these testify to the achievement of the purpose.
1.1	**Process performance attribute**
Level 2	**Managed Process** The process delivers work products of acceptable quality within defined timescales. Performance according to specified procedures is planned and tracked. Work products conform to specified standards and requirements. The primary distinction from the Performed Level is that the performance of the process is planned and managed and progressing towards a defined process.
2.1	**Performance management attribute**
2.2	**Work product management attribute**
Level 3	**Established Process** The process is performed and managed using a defined process based upon good software engineering principles. Individual implementations of the process use approved, tailored versions of standard, documented processes. The resources necessary to establish the process definition are also in place. The primary distinction from the Managed Level is that the process of the Established Level is planned and managed using a standard process.
3.1	**Process definition attribute**
3.2	**Process resource attribute**
Level 4	**Predictable Process** The defined process is performed consistently in practice within defined control limits, to achieve its goals. Detailed measures of performance are collected and analyzed. This leads to a quantitative understanding of process capability and an improved ability to predict performance. Performance is objectively managed. The quality of work products is quantitatively known. The primary distinction from the Established Level is that the defined process is quantitatively understood and controlled.
4.1	**Process measurement attribute**
4.2	**Process control attribute**
Level 5	**Optimising Process** Performance of the process is optimised to meet current and future business needs, and the process achieves repeatability in meeting its defined business goals. Quantitative process effectiveness and efficiency goals (targets) for performance are established, based on the business goals of the organisation. Continuous process monitoring against these goals is enabled by obtaining quantitative feedback and improvement is achieved by analysis of the results. Optimising a process involves piloting innovative ideas and technologies and changing non-effective processes to meet defined goals or objectives. The primary distinction from the Predictable Level is that the defined process and the standard process undergo continuous refinement and improvement, based on a quantitative understanding of the impact of changes to these processes.
5.1	**Process change attribute**
5.2	**Continuous improvement attribute**

Table 2: Overview of the capability levels and attributes.

The scope of an assessments is an Organisational Unit (OU) [24]. An OU deploys one or more processes that have a coherent process context and operates within a coherent set of business goals.

The characteristics that determine the coherent scope of activity - the process context - include the application domain, the size, the criticality, the complexity, and the quality characteristics of its products or services. An OU is typically part of a larger organisation, although in a small organisation the OU may be the whole organisation. An OU may be, for example, a specific project or set of (related) projects, a unit within an organisation focused on a specific life cycle phase (or phases), or a part of an organisation responsible for all aspects of a particular product or product set.

Rating & Designation	Description
Not Achieved - N	There is no evidence of achievement of the defined attribute.
Partially Achieved - P	There is some achievement of the defined attribute.
Largely Achieved - L	There is significant achievement of the defined attribute.
Fully Achieved - F	There is full achievement of the defined attribute.

Table 3: The four-point attribute rating scale.

3.2 Measuring SRA Process Capability

In ISO/IEC 15504, SRA is embodied in the *Develop software requirements* process. Requirements elicitation is covered by a different process, and therefore is not within the scope of our study. The purpose of the *Develop software requirements* process is to establish the requirements of the software component of the system. As a result of successful implementation of this process:

- the requirements allocated to software components of the system and their interfaces will be defined to match the customer's stated and implied needs;
- analyzed, correct and testable software requirements will be developed;
- the impact of software requirements on the operating environment will be understood;
- a relevant software release strategy will be developed that defines the priority for implementing software requirements;
- the software requirements will be approved and updated as needed;
- the software requirements will be communicated to all affected parties.

One of the ISO/IEC 15504 documents contains an exemplar assessment model (known as Part 5). This provides further details of how to rate the SRA process. Almost all of the assessments that were part of our study used Part 5 directly, and those that did not used models that are based on Part 5, therefore a discussion of the guidance for rating the SRA process in Part 5 is relevant here.

Basic practices that should exist to indicate that the purpose of the SRA process has been achieved are:

- **Specify software requirements.** Determine and analyse requirements of the software components of the system and document in a software requirements specification.
- **Determine operating environment impact.** Determine the interfaces between the software requirements and other components of the operating environment[15], and the impact that the requirements will have.
- **Evaluate requirements with customer.** Communicate the software requirements to the customer, and based on what is learned through this communication, revise if necessary.

[15] The operating environment includes tasks performed by or other systems used by the intended users of the software product.

- **Determine release strategy.** Prioritise the software requirements and map them to future releases of the software.
- **Update requirements for next iteration.** After completing an iteration of requirements, design, code, and test, use the feedback obtained from use to modify the requirements for the next iteration.
- **Communicate software requirements.** Establish communication mechanisms for dissemination of software requirements, and updates to requirements to all parties who will be using them.

For higher capabilities, a number of *Management Practices* have to be evaluated to determine the rating. For each of the attributes in levels 2 and 3, the management practices are summarised below. We do not consider levels above 3 because we do not include higher level ratings within our study.

3.2.1 Performance management attribute

The extent to which the execution of the process is managed to produce work products within stated time and resource requirements.

- In order to achieve this capability, a process needs to have time and resources requirements stated and produce work products within the stated requirements.
- The related Management Practices are:

Management practices
Identify resource requirements to enable planning and tracking of the process.
Plan the performance of the process by identifying the activities of the process and the allocated resources according to the requirements.
Implement the defined activities to achieve the purpose of the process.
Manage the execution of the activities to produce the work products within stated time and resource requirements.

3.2.2 Work product management attribute

The extent to which the execution of the process is managed to produce work products that are documented and controlled and that meet their functional and non-functional requirements, in line with the work product quality goals of the process.

- In order to achieve this capability, a process needs to have stated functional and non-functional requirements, including integrity, for work products and to produce work products that fulfil the stated requirements
- The related **Management Practices** are:

Management practices
Identify requirements for the integrity and quality of the work products.
Identify the activities needed to achieve the integrity and quality requirements for work products.
Manage the configuration of work products to ensure their integrity.
Manage the quality of work products to ensure that the work products meet their functional and non-functional requirements.

3.2.3 Process definition attribute

The extent to which the execution of the process uses a process definition based upon a standard process, that enables the process to contribute to the defined business goals of the organisation.

- In order to achieve this capability, a process needs to be executed according to a standard process definition that has been suitably tailored to the needs of the process instance. The standard process needs to be capable of supporting the stated business goals of the organisation.
- The related **Management Practices** are:

Management practices
Identify the standard process definition from those available in the organisation that is appropriate to the process purpose and the business goals of the organisation.
Tailor the standard process to obtain a defined process appropriated to the process context.
Implement the defined process to achieve the process purpose consistently, and repeatably, and support the defined business goal of the organisation.
Provide feedback into the standard process from experience of using the defined process.

3.2.4 Process resource attribute

The extent to which the execution of the process uses suitable skilled human resources and process infrastructure effectively to contribute to the defined business goals of the organisation.

- In order to achieve this capability, a process needs to have adequate human resources and process infrastructure available that fulfil stated needs to execute the defined process.
- The related **Management Practices** are:

Management practices
Define the human resource competencies required to support the implementation of the defined process.
Define process infrastructure requirements to support the implementation of the defined process.
Provide adequate skilled human resources meeting the defined competencies.
Provide adequate process infrastructure according to the defined needs of the process.

4 Research Method

4.1 Approaches to Evaluating Predictive Validity in Correlational Studies

		Measuring the Criterion		
		Questionnaire	**Measurement Program**	
Measuring Capability	**Questionnaire Assessment**	Q1	Q2	(low cost)
		Q3	Q4	(high cost)
		(across organisations)	(within one organisation)	

Table 4: Different correlational approaches for evaluating predictive validity.

Correlational approaches to evaluating the predictive validity of a process capability measure can be classified by the manner in which the variables are measured. Table 4 shows a classification of approaches. The columns indicate the manner in which the criterion is measured. The rows indicate the manner in which the process capability is measured. The criterion can be measured using a questionnaire whereby data on the perceptions of experts are collected. It can also be measured through a measurement program. For example, if our criterion is defect density of delivered software products, then this could be measured through an established measurement program that collects data from defects found in the field. Process capability can also be measured through a questionnaire whereby data on the perceptions of experts on the capability of their processes are collected. Alternatively, actual assessments can be performed, which are a more rigorous form of measurement[16].

Conducting a study where capability is measured through an assessment, and the criterion is measured through a measurement program provides for more rigour.[17] However, as also indicated in Table 4, assessments are more costly and studies that utilise data from a measurement program will almost always be conducted within a single organisation, hence reducing the generalisability of their results. Therefore, the selection of a quadrant in Table 4 is a tradeoff amongst cost, measurement rigour, and generalisability.

Many previous studies that evaluated the relationship between process capability (or organisational maturity) and the performance of projects tended to be in quadrant Q1. For example, [34][16][10]. These studies have the advantage that they can be conducted across multiple projects and across multiple organisations, and hence can produce more generalisable conclusions.

---

[16] "More rigorous" is intended to mean with greater reliability and construct validity.

[17] A difficulty with this approach is that the majority of organisations do not collect objective process and product data (e.g., on defect levels, or even keep accurate effort records). Organisations following the benchmarking paradigm do not necessarily have measurement programs in place to provide the necessary data. Primarily organisations that have made improvements and reached a reasonable level of process capability will have the actual objective data to demonstrate improvements (in productivity, quality, or return on investment). This assertion is supported by the results in [7] where, in general, it was found that organisations at lower SW-CMM maturity levels are less likely to collect quality data (such as the number of development defects). Also, the same authors found that organisations tend to collect more data as their CMM maturity levels rise. It was also reported in another survey [74] that for 300 measurement programs started since 1980, less than 75 were considered successful in 1990, indicating a high mortality rate for measurement programs. This high mortality rate indicates that it may be difficult right now to find many organisations that have implemented measurement programs.

This means that organisations or projects with low process capability would have to be excluded from a correlational study. Such an exclusion would reduce the variation in the performance measure, and thus reduce (artificially) the validity coefficients. Therefore, correlational studies that utilise objective performance measures are inherently in greater danger of not finding significant results, especially if the data is collected across multiple organisations.

Another difficulty is to ensure that the performance measures are defined and measured consistently across multiple organisations. For example, the definition of a defect would be the same in measures of defect density.

A more recent study evaluated the relationship between questionnaire responses on implementation of the SW-CMM KPA's and defect density [50], and this would be placed in quadrant Q2. However, this study was conducted across multiple projects within a single organisation, reducing its generalisability compared with studies conducted across multiple organisations.

Our current study can be placed in quadrant Q3 since we use process capability measures from actual assessments, and questionnaires for evaluating project performance. This retains the advantage of studies in quadrant Q1 since it is conducted across multiple projects in multiple organisations, but utilises a more rigorous measure of process capability. Similarly, the study of Jones can be considered to be in this quadrant [45][46]. [18]

Studies in quadrant Q4 are likely to have the same limitations as studies in quadrant Q2: being conducted across multiple projects within the same organisation. For instance, the study of McGarry et al was conducted within a single company [59], and the AFIT study was conducted with contractors of the Air Force [28][53].

Therefore, the different types of studies that can be conducted in practice have different advantages and disadvantages, and predictive validity studies have been conducted in the past that populate all four quadrants. It is reasonable then to encourage studies in all four quadrants. Consistency in the results across correlational studies that use the four approaches would increase the weight of evidence supporting the predictive validity hypothesis.

4.2 Source of Data

The data that was used for this study was obtained from Phase 2 of the SPICE Trials. During the trials, organisations contribute their assessment ratings data to an international trials database located in Australia, and also fill up a series of questionnaires after each assessment. The questionnaires collect information about the organisation and about the assessment. There is a network of SPICE Trials co-ordinators around the world who interact directly with the assessors and the organisations conducting the assessments. This interaction involves ensuring that assessors are qualified, making questionnaires available, answering queries about the questionnaires, and following up to ensure the timely collection of data.

Region	Number of Assessments
Canada	1
Europe	24
North Asia Pacific	10
South Asia Pacific	34
USA	1

Table 5: Distribution of assessments by region.

[18] Since it is difficult to find low maturity organisations with objective data on effort and defect levels, and since there are few high maturity organisations, Jones' data relies on the reconstruction of, at least, effort data from memory, as noted in [44]: "The SPR approach is to ask the project team to reconstruct the missing elements from memory." The rationale for that is stated as "the alternative is to have null data for many important topics, and that would be far worse." The general approach is to show staff a set of standard activities, and then ask them questions such as which ones they used and whether they put in any unpaid overtime during the performance of these activities. For defect levels, the general approach is to do a matching between companies that do not measure their defects with similar companies that do measure, and then extrapolate for those that don't measure. It should be noted that SPR does have a large data base of project and organisational data, which makes this kind of matching defensible. However, since at least some of the criterion measures are not collected from measurement programs, we place this study in the same category as those that utilise questionnaires.

Region	Number of OUs
Canada	1
Europe	23
North Asia Pacific	3
South Asia Pacific	16
USA	1

Table 6: Distributiuon of assessed OUs by region.

At the time of writing a total of 70 assessments had been conducted. The distribution of assessments by region is given in Table 5.[19] In total 691 process instances were assessed. Since more than one assessment may have occurred in a particular OU (e.g., multiple assessments each one looking at a different set of processes), a total of 44 OUs were assessed. Their distribution by region is given in Table 6.

4.3 Data Analysis

4.3.1 Measurement

A previous study had identified that the capability scale of ISO/IEC 15504 is two dimensional [20]. The first dimension, which was termed "Process Implementation", consists of the first three levels. The second dimension, which was termed "Quantitative Process Management", consists of levels 4 and 5. It was also found that these two dimensions are congruent with the manner in which assessments are conducted in practice: either only the "Process Implementation" dimension is rated or both dimensions are rated (recall that it is not required to rate at all five levels in an ISO/IEC 15504 assessment).

In our data set, 36% of the SRA processes were not rated on the "Quantitative Process Management" dimension. If we exclude all processes with this rating missing then we lose a substantial proportion of our observations. Therefore, we limit ourselves in the current study to the first dimension only.

To construct a single measure of "Process Implementation" we code an 'F' rating as 4, down to a 1 for an 'N' rating. Subsequently, we construct an unweighted sum of the attributes at the first three levels of the capability scale. This is a common approach for the construction of summated rating scales [60].

The performance measures were collected through a questionnaire. The respondent to the questionnaire was the sponsor of the assessment, who should be knowledgeable about the projects that were assessed. In cases where the sponsor was not able to respond, s/he delegated the task to a project manager or senior technical person who completed the questionnaire. The responses were coded such that the "Excellent" response category is 1, down to the "Poor" response category which was coded 4. The "Don't Know" responses were treated as missing values. The implication of this coding scheme is that all investigated relationships are hypothesised to be negative.

4.3.2 Evaluating the Relationships

We follow a two staged analysis procedure. During the first stage we determine whether the association between "Process Implementation" of the SRA process and each of the performance measures is "clinically significant". This is done using the Pearson product moment correlation coefficient. This means that it has a magnitude that is sufficiently large. If it is, then we test the statistical significance of the association. The logic of this is explained below.

It is known that with a sufficiently large sample size even very small associations can be statistically significant. Therefore, it is also of import to consider the magnitude of a relationship to determine whether

[19] Within the SPICE Trials, assessments are coordinated within each of the five regions shown in the figures above.

it is large. Cohen has provided some general guidelines for interpreting the magnitude of the correlation coefficient [13]. We consider "medium" sized (i.e., r =0.3) correlations as the minimal magnitude that is worthy of consideration. The logic behind this choice is that of elimination. If we take "small" association (i.e., r = 0.1) as the minimal worthy of consideration we may be being too liberal and giving credit to weak associations that are not congruent with the broad claims made for the predictive validity of assessment scores. Using a "large" association (i.e., r = 0.5) as the minimal value worthy of consideration may place a too high expectation on the predictive validity of assessment scores; recall that many other factors are expected to influence the success of a software project apart from the capability of the SRA process.

For statistical significance testing, we perform an ordinary least squares regression:

$$PERF = \hat{a} + \left(\hat{Q} \times CAP\right) \qquad \textbf{Eqn. 9}$$

where $PERF$ is the performance measure according to Table 1 and CAP is the "Process Implementation" dimension of process capability. We test whether the $\hat{Q}$ regression coefficient is different from zero. If there is sufficient evidence that it is (we define sufficient evidence in Section 4.3.4), then we claim that CAP is associated with $PERF$. The above model is constructed separately for each of the performance measures.

4.3.3 Scale Type Assumption

According to some authors, one of the assumptions of the OLS regression model (e.g., see [4]) is that all the variables should be measured at least on an interval scale. This assumption is based on the mapping originally developed by Stevens [88] between scale types and "permissible" statistical procedures. In our context, this raises two questions. First, what are the levels of our measurement scales? Second, to what extent can the violation of this assumption have an impact on our results?

The scaling model that is used in the measurement of the process capability construct is the summative model [60]. This consists of a number of subjective measures on a 4-point scale that are summed up to produce an overall measure of the construct. Some authors state that summative scaling produces interval level measurement scales [60], while others argue that this leads to ordinal level scales [32]. In general, however, our process capability is expected to occupy the grey region between ordinal and interval level measurement.

Our criterion measures utilised a single item each. In practice, single item measures are treated as if they are interval in many instances. For example, in the construction and empirical evaluation of the User Information Satisfaction instrument, inter-item correlations and principal components analysis are commonly performed [43].

It is also useful to note a study by Spector [86] that indicated that whether scales used have equal or unequal intervals does not actually make a practical difference. In particular, the mean of responses from using scales of the two types do not exhibit significant differences, and that the test-retest reliabilities (i.e., consistency of questionnaire responses when administered twice over a period of time) of both types of scales are both high and very similar. He contends, however, that scales with unequal intervals are more difficult to use, but that respondents conceptually adjust for this.

Given the proscriptive nature of Stevens' mapping, the permissible statistics for scales that do not reach an interval level are distribution-free (or nonparametric) methods (as opposed to parametric methods, of which multiple regression is one) [80]. Such a broad proscription is viewed by Nunnally as being "narrow" and would exclude much useful research [62]. Furthermore, studies that investigated the effect of data transformations on the conclusions drawn from parametric methods (e.g., F ratios and t tests) found little evidence supporting the proscriptive viewpoint [52][51][1]. Suffice it to say that the issue of the validity of the above proscription is, at best, debatable. As noted by many authors, including Stevens himself, the basic point is that of pragmatism: useful research can still be conducted even if, strictly speaking, the proscriptions are violated [88][4][33][91]. A detailed discussion of this point and the literature that supports our argument is given in [5].

4.3.4 Multiple Hypothesis Testing

Since we are performing multiple hypotheses testing (i.e., for each one of the regression models), it is plausible that many $\hat{Q}$ regression coefficients will be found to be statistically significant since the more null hypothesis tests that one performs, the greater the probability of finding statistically significant results by chance. We therefore use a Bonferonni adjusted alpha level when performing significance testing [65]. We set our overall alpha level to be 0.1.

4.3.5 Organisation Size Context

It was noted earlier that the relationships may be of different magnitudes for small vs. large organisations. We therefore perform the analysis separately for small and large organisations. Our definition of size is the number of IT staff within the OU. We dichotomise this IT staff size into SMALL and LARGE organisations, whereby small is equal to or less than 50 IT staff. This is the same definition of small organisations that has been used in a European project that is providing process improvement guidance for small organisations [87].

4.3.6 Reliability of Measures

It is known that lack of reliability in measurement can attenuate bivariate relationships [62]. It is therefore important to evaluate the reliability of our subjective measures, and if applicable, make corrections to the correlation coefficient that take into account reliability.

In another related scientific discipline, namely Management Information Systems (MIS), researchers tend to report the Cronbach alpha coefficient [14] most frequently [89]. Also, it is considered by some researchers to be the most important reliability estimation approach [79]. This coefficient evaluates a certain type of reliability called internal consistency, and has been used in the past to evaluate the reliability of the ISO/IEC 15504 capability scale [20][31]. We also calculate the Cronbach alpha coefficient for the SRA process capability measure.

In our study we do not incorporate corrections for attenuation due to less than perfect reliability, however. As suggested in [62], it is preferable to use the unattenuated correlation coefficient since this reflects the predictive validity of the process capability measure that will be used in actual practice (i.e., in practice it will have less than perfect reliability).

4.3.7 Multiple Imputation

In the performance measures that we used (see Table 1) there were some missing values. Missing values are due to respondents not providing an answer on all or some of the performance questions, or they selected the "Don't Know" response category. Ignoring the missing values and only analysing the completed data subset can provide misleading results [58]. We therefore employ the method of multiple imputation to fill in the missing values repeatedly. Multiple imputation is a preferred approach to handling missing data problems in that it provides for proper estimates of parameters and their standard errors.

The basic idea of multiple imputation is that one generates a vector of size M for each value that is missing. Therefore an $n_{mis} \times M$ matrix is constructed, where n_{mis} is the number of missing values. Each column of this matrix is used to construct a complete data set, hence one ends up with M complete data sets. Each of these data sets can be analysed using complete-data analysis methods. The M analyses are then combined into one final result. Typically a value for M of 3 is used, and this provides for valid inference [73]. Although, to err on the conservative side, some studies have utilised an M of 5 [90], which is the value that we use.

For our analysis the two parameters of interest are the correlation coefficient, r, and the $\hat{Q}$ parameter of the regression model. Furthermore, we are interested in the standard error of $\hat{Q}$, which we shall denote as $\sqrt{U}$, in order to test the null hypothesis that it is equal to zero. After calculating these values for each of the 5 data sets, they can be combined to give an overall r value, $\bar{r}$, an overall value for $\hat{Q}$, $\overline{Q}$, and

its standard error $\sqrt{T}$. Procedures for performing this computation are detailed in [70], and summarised in [73]. In Section 7 we describe the multiple imputation approach in general, its rationale, and how we operationalised it for our specific study.

4.3.8 Summary of Data Analysis Method

The following steps summarize our data analysis method:

- Caluclate the Cronbach alpha reliability coefficient.
- Generate 5 complete data sets using multiple imputation.
- For each of the imputed data sets, build a regression model as defined in Eqn. 9 for each OU size.
- Combine the results of the five regression models into one result.
- Interpret the results according to the guidelines in Section 4.3.2, and using the Bonferonni adjustment.

5 Results

5.1 Description of Projects and Assessments

In this section we present some descriptive statistics on the projects that were assessed, and on the assessments themselves. In the SPICE Phase 2 trials, a total of 44 organisations participated. Their business sector distribution is summarised in Figure 4. As can be seen, the most frequently occurring categories are Defence, IT Products and Services, and Software Development organisations.

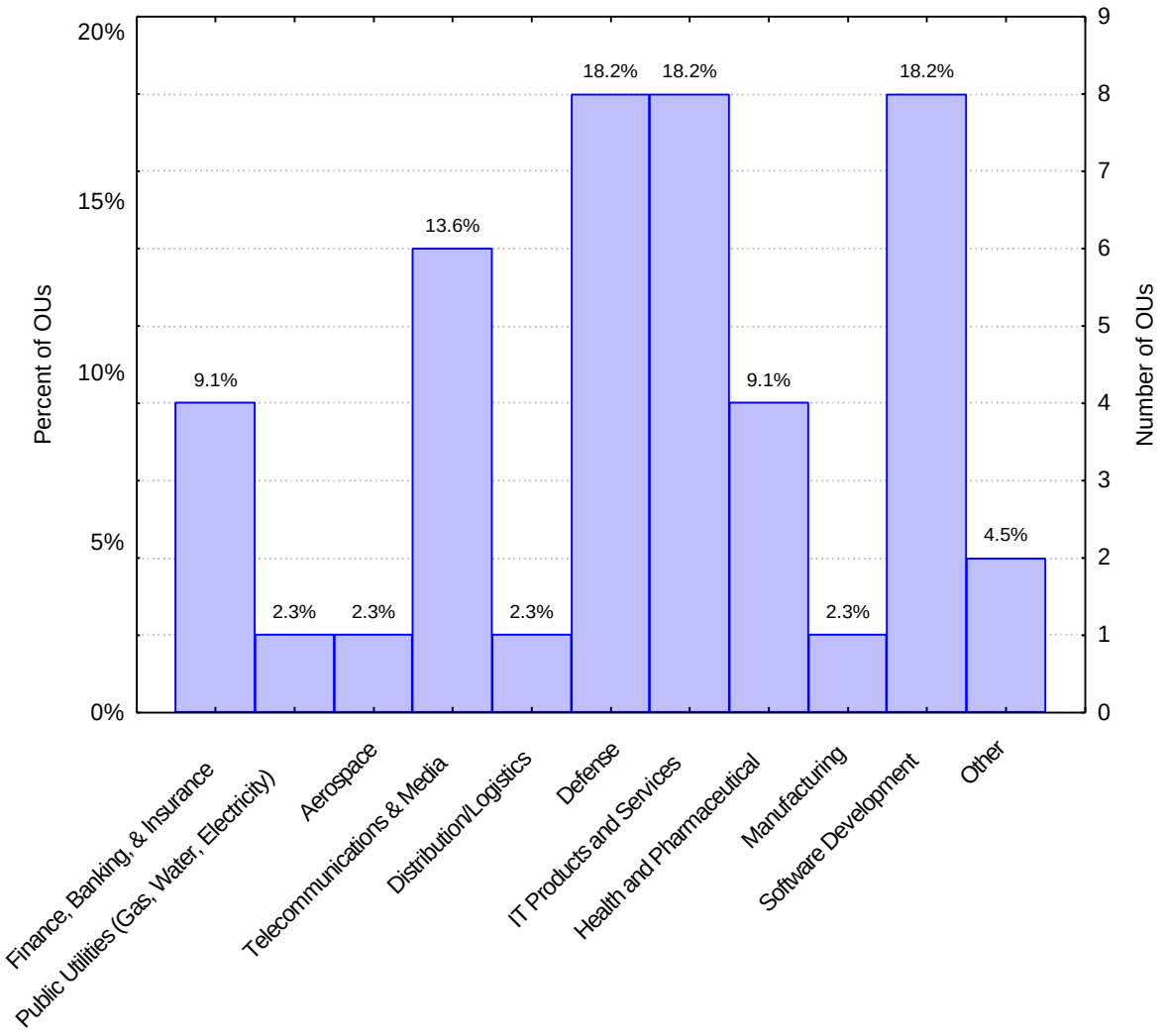

Figure 4: Business sector of all organisations that took part in SPICE Trials Phase 2 assessments (n=44).

Figure 5 shows the distributions for those 29 organisations that assessed the SRA process. While the three most frequent categories in Figure 4 are still the most frequent in Figure 5, organisations in the Finance, Banking, and Insurance business sector also tend to have a relatively high presence in this subset.

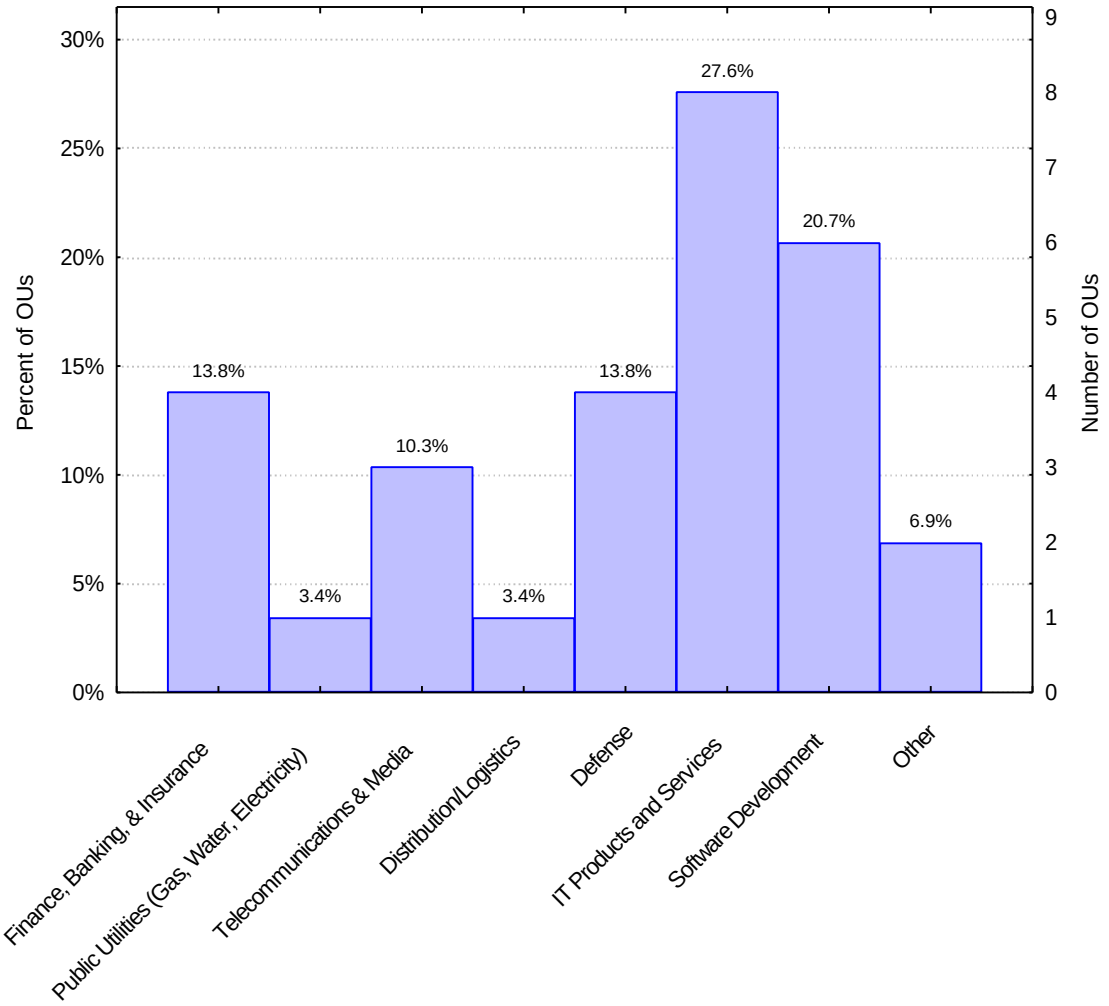

Figure 5: Business sector of all organisations that assessed the requirements engineering process (n=29).

Of the 29 OUs, they were distributed by country as follows: Australia (10), Singapore (1), Canada (1), Germany (2), Italy (1), France (2), Spain (2), Turkey (1), Luxemburg (3), South Africa (4), Hungary (1), and Japan (1). Of these 14 were not ISO 9001 registered, and 15 were ISO 9001 registered.

In total, the SRA process from 56 projects within the 29 OUs assessed their SRA process. Figure 6 shows the variation in the number of projects that were assessed in each OU. The median value is two projects assessed within a single OU, although clearly some OUs had up to eight projects assessed. Since more than one project can be within the scope of a single assessment, 35 different assessments were conducted.

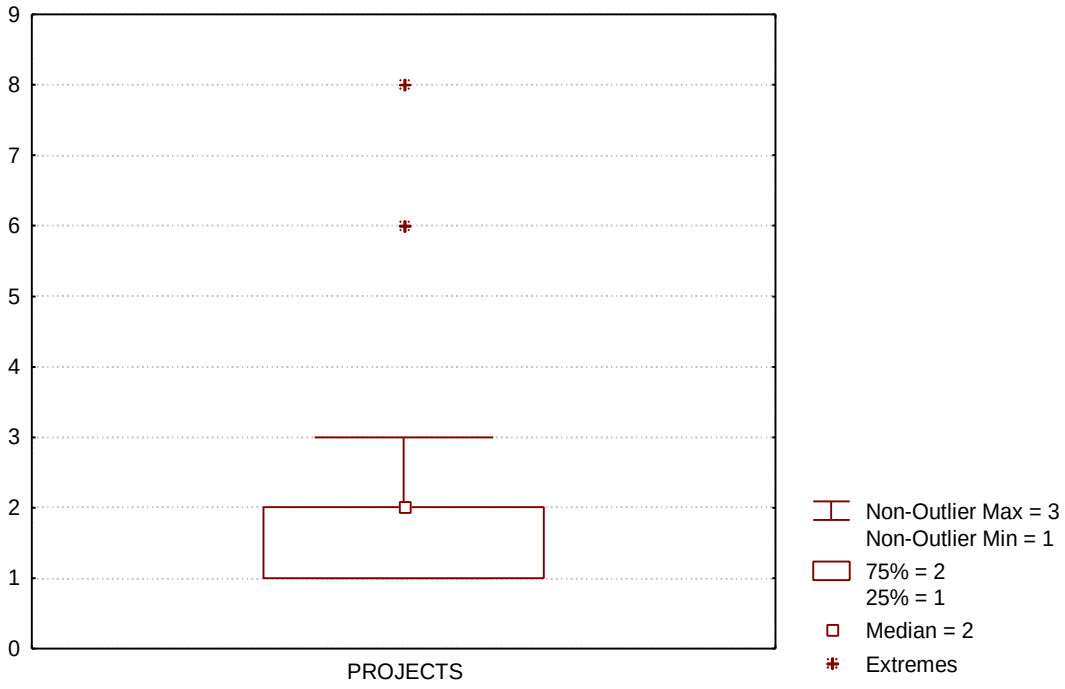

Figure 6: Variation in the number of projects that were assessed in each OU.

The distribution of peak staff load for the 56 projects is shown in Figure 7. The median value is 6 staff at peak time, although some projects had a peak of 80 staff, and some as low as one staff working on them.

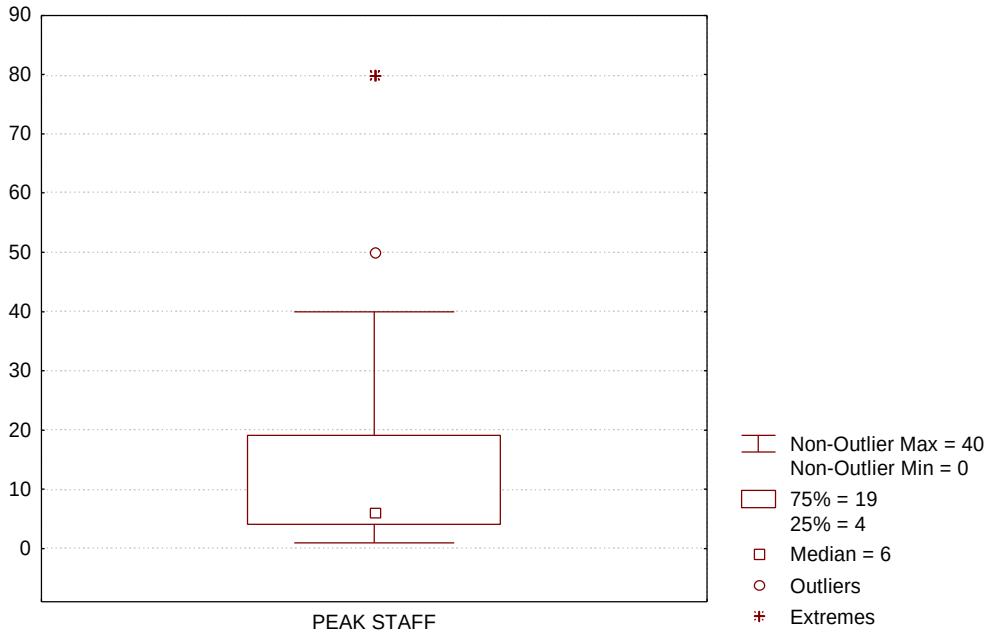

Figure 7: Variation in the peak staff load for the assessed projects. Note that this is different from the number of IT staff in the whole of the OU.

In Figure 8 we can see the variation in the two measures of process capability. For D2 (Quantitative Process Management), there is little variation. If we only consider the projects that assessed the SRA

process along that dimension, the maximum obtained value was 10, which is exacty half the maximum possible value. The median is 4. This indicates that very few projects actually achieve high capability along the D2 dimension.

The minimum score of 6 on D1 (Process Implementation) is indicative of the fact that all projects did perform the SRA activities that were defined earlier to some level (i.e., none received the "Not Achieved" rating on the Performance attribute). Also, the maxuimum value of 20 indicates that there were some projects that met all of the requirements of the "Process Implementation" dimension for the SRA process.

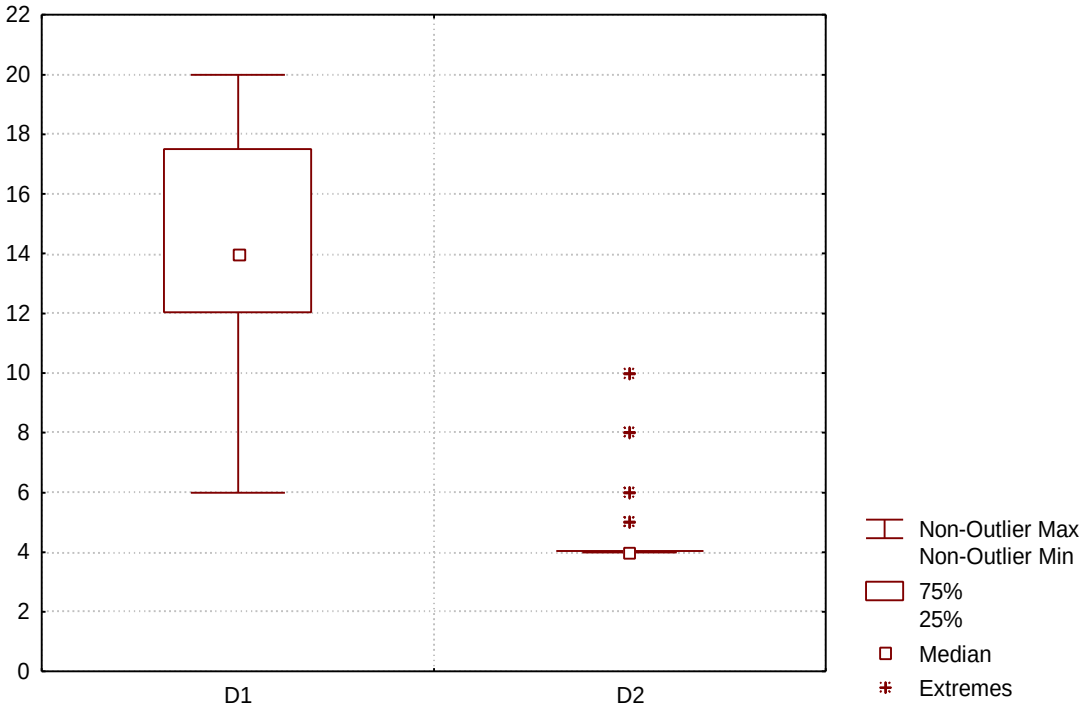

Figure 8: Variation in the measures of the two dimensions of process capability (denoted D1 and D2). For the second dimension it is assumed that processes that were not rated would receive a rating of N (Not Achieved) if they would have been rated. This was done to ensure that the sample size for both dimensions was the same. Note that it is common practice not to rate the higher levels if there is strong *a priori* belief that the ratings will be N.

Figure 9 shows the variation along the D1 dimension for the two sizes that were considered in our study. While the min-max ranges are the same, larger OUs tend to have higher capability than small OUs. Although the difference is not marked.

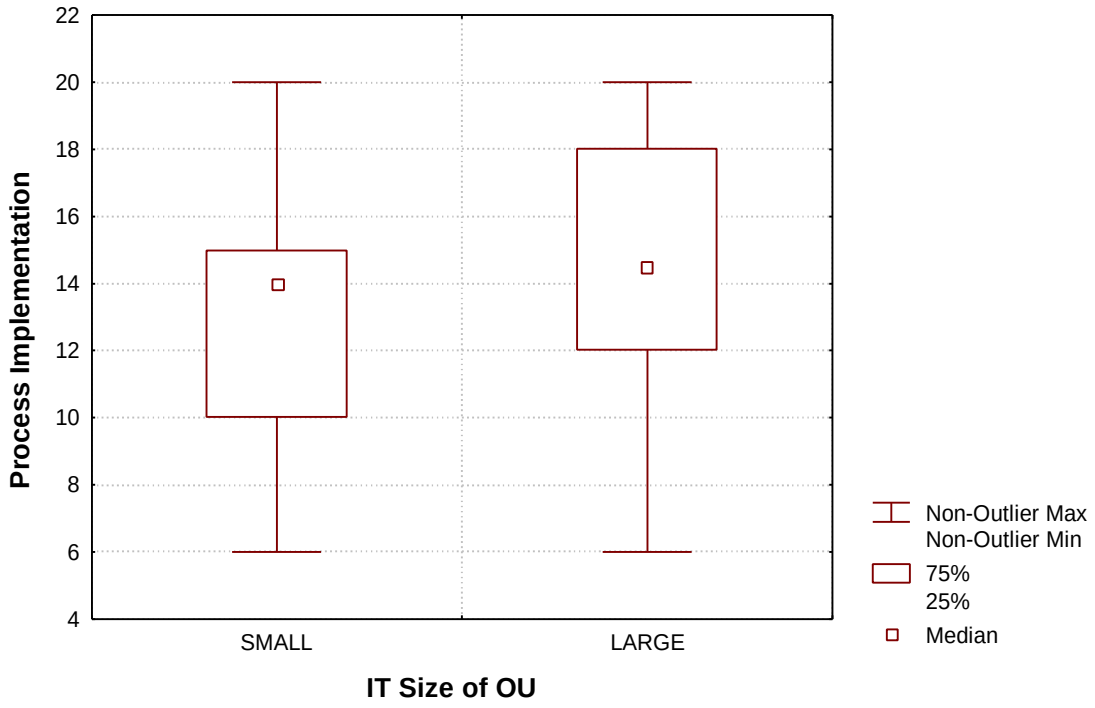

Figure 9: The variation in the D1 dimension for the different OU sizes that were considered in our study (IT staff). In total, there were 22 projects in "small" organisations, and 34 projects in "large" organisations.

5.2 Reliability of the SRA Process Capability Measure

The Cronbach alpha coefficient for the "Process Implementation" variable was found to be 0.84 when evaluated on only the SRA process. For the purposes of our study, this can be considered sufficiently large [62].[20]

5.3 Affects of SRA Process Capability

Table 7 shows the results for small organisations, and Table 8 for large organisations. The tables show the $\hat{Q}$ coefficient and its standard error for each imputed complete data set. The combined results include the average correlation coefficient across the complete data sets ($\bar{r}$), and the average $\hat{Q}$ coefficient ($\overline{Q}$) and its multiply imputed standard error $\sqrt{T}$.

From Table 7 we can see that none of the $\bar{r}$ values reach a value of 0.3, although the value of the REQUIREMENTS dependent variable does approach this threshold. We therefore conclude that there is no evidence suggesting that increases in "Process Implementation" for the SRA process is related to any of our project performance measures.

For large organisations, we can see from Table 8 that the value for the PRODUCTIVITY and MORALE dependent variables are both above our threshold 0.3. We therefore proceed to statistical testing. Since we set the overall alpha level at 0.1, we test each $\hat{Q}$ coefficient at the Bonferroni adjusted alpha level of 0.05. For this we found that only the PRODUCTIVITY dependent variable was significantly related with the "Process Implementation" capability of the SRA process. This means that improvements in SRA

[20] Nunnally and Bernstein [62] recommend that a coefficient of 0.8 is a minimal threshold for applied research settings, and a minimal threshold of 0.7 for basic research settings.

process capability are associated with a reduction in the cost of software projects. This can be interpreted to be attributable to less rework during later phases of the project.

Not finding a relationship with BUDGET and SCHEDULE is perhaps not surprising. Ability to meet schedule and budget commitments depends on making realistic commitments to start off with. The capability of the SRA process by itself will not ensure that commitments are realistic. Other factors, such as the quality of cost and schedule estimation practices, for example, will have a substantial impact on BUDGET and SCHEDULE. Therefore, SRA process capability by itself is not sufficient to explain much variation in these two dependent variables.

The lack of relationship with customer satisfaction follows from not finding a relationship with some of the other variables. For example, if organisations that have a high SRA process capability still do not meet budget and schedule commitments and do not satisfy specified requirements then it is not surprising that this will also not result in greater customer satisfaction.

Inability to find a relationship with satisfaction of requirements may be a consequence of this outcome being influenced by other factors later in the life cycle. For example, if a project's verification processes are weak, then there will be no feedback to the project on the extent to which requirements have been satisfied. Therefore, SRA process capability may be an insufficient condition for ensuring requirements satisfaction.

Even though MORALE has a strong relationship with SRA process capability, it does not attain statistical significance. One would expect that better ability to manage requirements and reduce rework would result in improved staff morale and job satisfaction. The fact that it does not may be due to a number of reasons:

- The Bonferroni procedure is conservative.
- Job satisfaction and morale will likely be affected by a multitude of other variables, some of which are unrelated to processes at all (for example, compensation and working environment). Furthermore, it can be argued that improved process capability on a single process will be less influential if capability on other processes is weak.

	Results from Repeated Imputations										**Combined Results**		
	Imputation 1		**Imputation 2**		**Imputation 3**		**Imputation 4**		**Imputation 5**				
	$\hat{Q}_1$	$\sqrt{U_1}$	$\hat{Q}_2$	$\sqrt{U_2}$	$\hat{Q}_3$	$\sqrt{U_3}$	$\hat{Q}_4$	$\sqrt{U_4}$	$\hat{Q}_5$	$\sqrt{U_5}$	$\bar{r}$	$\bar{Q}$	$\sqrt{T}$
BUDGET	0.011	0.045	-0.013	0.042	-0.015	0.041	0.021	0.052	-0.015	0.043	-0.017	-0.002	0.0484
SCHEDULE	-0.055	0.044	-0.06	0.047	-0.095	0.044	-0.066	0.052	-0.028	0.051	-0.274	-0.060	0.0544
CUSTOMER	0.045	0.033	0.013	0.033	0.028	0.032	0.037	0.036	0.039	0.031	0.2134	0.032	0.0357
REQUIREMENTS	0.028	0.029	0.037	0.029	0.043	0.033	0.053	0.033	0.054	0.03	0.296	0.043	0.0331
PRODUCTIVITY	0.011	0.032	-0.012	0.03	-0.012	0.025	-0.012	0.025	0.019	0.032	-0.021	-0.001	0.0333
MORALE	0.046	0.039	0.002	0.034	0.044	0.036	0.044	0.036	0.009	0.035	0.1726	0.029	0.0431

Table 7: Repeated imputation results and combined results for small organisations.

	Results from Repeated Imputations										Combined Results		
	Imputation 1		Imputation 2		Imputation 3		Imputation 4		Imputation 5				
	$\hat{Q}_1$	$\sqrt{U_1}$	$\hat{Q}_2$	$\sqrt{U_2}$	$\hat{Q}_3$	$\sqrt{U_3}$	$\hat{Q}_4$	$\sqrt{U_4}$	$\hat{Q}_5$	$\sqrt{U_5}$	$\bar{r}$	$\overline{Q}$	$\sqrt{T}$
BUDGET	0.007	0.049	-0.014	0.049	-0.074	0.049	0.074	0.043	0.011	0.052	0.009	0.0008	0.0758
SCHEDULE	0.061	0.049	0.068	0.05	0.106	0.049	0.09	0.049	0.002	0.049	0.2242	0.0654	0.0656
CUSTOMER	-0.011	0.031	-0.023	0.031	0.016	0.033	-0.043	0.03	0.009	0.037	-0.062	-0.01	0.0417
REQUIREMENTS	0.036	0.041	0.054	0.038	0.059	0.037	0.086	0.038	0.052	0.037	0.2554	0.0574	0.0430
PRODUCTIVITY	-0.162	0.034	-0.151	0.034	-0.162	0.038	-0.162	0.034	-0.162	0.036	**-0.625**	**-0.159**	0.0356
MORALE	-0.052	0.028	-0.083	0.028	-0.029	0.029	-0.038	0.029	-0.063	0.027	**-0.312**	-0.053	0.0365

Table 8: Repeated imputation results and combined results for large organisations.

Our results indicate that increased process capability on the SRA process is related to a reduction in project costs for large software organisations. This result by itself is encouraging since data collected by SPR indicate that 100% of top management consider the reduction in software development costs as one of their goals [37]. In another survey by the Meta Group, increasing productivity was ranked higher than project management, improving software quality, and the implementation of metrics, and is ranked as the seventh top management priority in the US [75]. To our knowledge, this is the first study to evaluate the predictive validity of the SRA process capability using an internationally standardised measure of capability.

In addition to this particular result, we can tentatively draw a number of conclusions:

- Evaluation of predictive validity of individual processes is more informative than the evaluation of aggregate measures of capability (i.e., aggregating across many processes). This point was made in the past [26], with the argument being that the capability of an individual process is likely not to be related with *all* project performance measures. This was certainly true in our case.
- There seems to be a substantial difference between the effects of SRA process capability on project performance for small and large organisations. Our study found that SRA process capability was not related to any project performance measures for small organisations.

5.4 Limitations

One potential limitation of our results concerns their generalisability. Specifically, the extent to which our findings can be generalised to assessments that are not based on the emerging ISO/IEC 15504 International Standard. The emerging ISO/IEC 15504 International Standard defines requirements on assessments. Assessments that satisfy the requirements are claimed to be compliant. Based on public statements that have been made thus far, it is expected that some of the more popular assessment models and methods will be consistent with the emerging ISO/IEC 15504 International Standard. For example, Bootstrap version 3.0 claims compliance with ISO/IEC 15504 [3], and the future CMMI product suite is expected to be consistent and compatible [83]. The assessments from which we obtained our data are also considered to be compliant. The extent to which our results, obtained from a subset of compliant assessments, can be generalised to all compliant assessments is an empirical question and can be investigated through replications of our study. The logic of replications leading to generalisable results is presented in [56].

Another limitation of our study is that the hypothesized model that we tested did not account for all possible factors that may have an impact on performance. Future work should extend this model accordingly.

6 Conclusions

In this paper we have presented an empirical study that evaluated the predictive validity of the ISO/IEC 15504 measure of software requirements analysis process capability. Predictive validity is a basic premise of all software process assessments that produce quantitative results. We first demonstrated that no previous studies have evaluated the predictive validity of this process, and then described our study in detail. Our results indicate that higher SRA process capability is related with increased productivity in software projects for large organisations. No evidence of predictive validity was found for small organisations.

The results indicate that improving the SRA process may potentially lead to improvements in the productivity of software projects in large organisations. It is by no means claimed that SRA process capability is the only factor that is associated with productivity. Only that a relatively strong association has been found during our study, suggesting that the SRA process ought to be considered as a target process for assessment and improvement if the objective of the organisation is to improve its productivity.

It is important to emphasise that studies such as this ought to be replicated to provide further confirmatory evidence as to the predictive validity of SRA process capability. It is known in scientific pursuits that there exists a "file drawer problem" [68]. This problem occurs when there is a reluctance by journal editors to publish, and hence a reluctance by researchers to submit, research results that do not show statistically significant relationships. One can even claim that with the large vested interest in the software process assessment community, reports that do not demonstrate the efficacy of a particular approach or model may be buried and not submitted for publication. Therefore, published works are considered to be a biased sample of the predictive validity studies that are actually conducted. However, by combining the results from a large number of replications that show significant relationships, one can assess the number of studies showing no significant relationships that would have to be published before our overall conclusion of there being a significant relationship is put into doubt [68]. This assessment would allow the community to place realistic confidence in the results of published predictive validity studies.

Future work should focus on evaluating other processes defined in ISO/IEC 15504. Given that ISO/IEC 15504 is in the final stages of becoming a full International Standard, the existence of evidence to support its use would give the software engineering community confidence in its usage. Furthermore, improvements in the measurement of variables would help strengthen the conclusions that can be drawn from validation studies. For example, by considering actual defect density and productivity values rather than values obtained through questionnaires.

7 Appendix: Multiple Imputation Method

In this appendix we describe the approach that we used for imputing missing values on the performance variable, and also how we operationalise it in our specific study. It should be noted that, to our knowledge, multiple imputation techniques have not been employed thus far in software engineering empirical research, where the common practice has been to ignore observations with missing values.

7.1 Notation

We first present some notation to facilitate explaining the imputation method. Let the raw data matrix have i rows (indexing the cases) and j columns (indexing the variables), where $i = 1 \ldots n$ and $j = 1 \ldots q$. Some of the cells in this matrix may be unobserved (i.e., missing values). We assume that there is only one outcome variable of interest for imputation (this is also the context of our study since we deal with each dependent variable seperately), and let y_i denote its value for the i^{th} case. Let $Y = (Y_{mis}, Y_{obs})$, where Y_{mis} denotes the missing values and Y_{obs} denotes the observed values on that variable. Furthermore, let X be a scaler or vector of covariates that are fully observed for every i. These may be background variables, which in our case were the size of an organisation in IT staff and whether the organisation was ISO 9001 registered, and other covariates that are related to the outcome

variable, which in our case was the process capability measure (i.e., “process implementation” as defined in the main body of the text).

Let the parameter of interest in the study be denoted by Q. We assume that Q is scaler since this is congruent with our context. For example, let Q be a regression coefficient. We wish to estimate $\hat{Q}$ with associated variance U from our sample.

7.2 Ignorable Models

Models underlying the method of imputation can be classified as assuming that the reasons for the missing data are either ignorable or nonignorable. Rubin [70] defines this formally. However, here it will suffice to convey the concepts, following [71].

Ignorable reasons for the missing data imply that a nonrespondent is only randomly different from a respondent with the same value of X. Nonignorable reasons for missing data imply that, even though respondents and nonrespondents have the same value of X, there will be a systematic difference in their values of Y. An example of a nonignorable response mechanism in the context of process assessments that use a model such as that of ISO/IEC 15504 is when organisations assess a particular process because it is perceived to be weak and important for their business. In such a case, processes for which there are capability ratings are likely to have lower capability than other processes that are not assessed.

In general, most imputation methods assume ignorable nonresponse [78] (although, it is possible to perform, for example, multiple imputation, with a nonignorable nonresponse mechanism). In the analysis presented in this report there is no a priori reason to suspect that respondents and nonrespondents will differ systematically in the values of the outcome variable, and therefore we assume ignorable nonresponse.

7.3 Overall Multiple Imputation Process

The overall multiple imputation process is shown in Figure 10. Each of these tasks is described below. It should be noted that the description of these tasks is done from a Bayesian perspective.

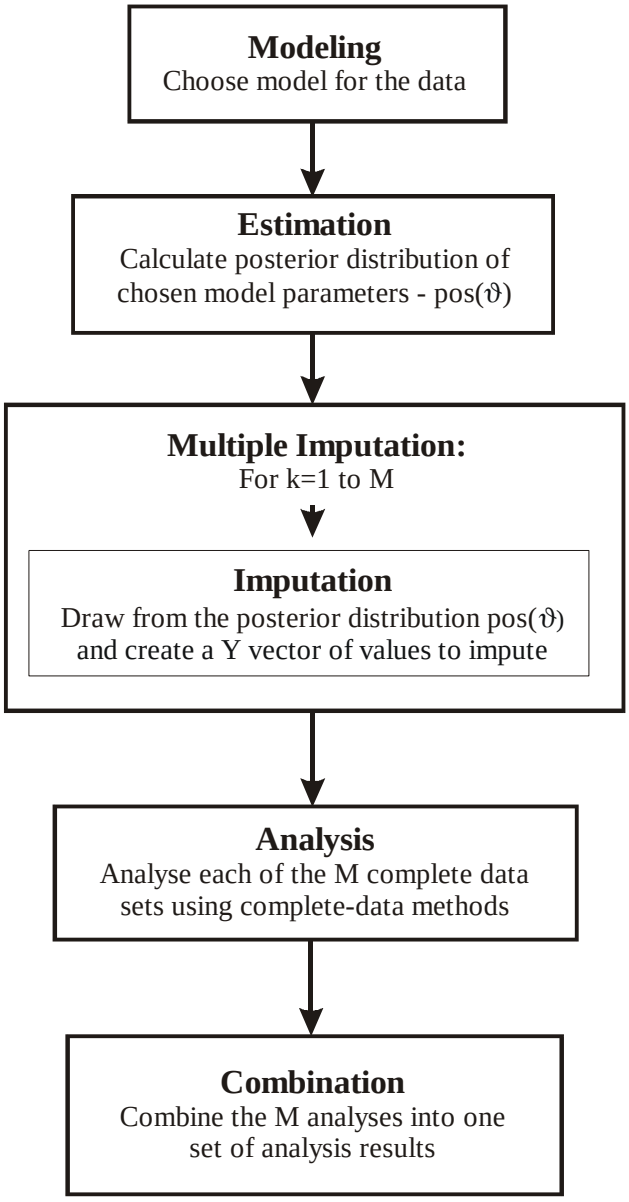

Figure 10: Schematic showing the tasks involved in multiple imputation.

7.4 Modelling Task

The objective of the modelling task is to specify a model $f_{Y|X}\left(Y_i \middle| X_i, q_{Y|X}\right)$ using the observed data only where $q_{Y|X}$ are the model parameters. For example, consider the situation where we define an ordinary least squares regression model that is constructed using the observed values of Y and the predictor variables are the covariates X, then $q_{Y|X} = (b, s^2)$ are the vector of the regression parameters and the variance of the error term respectively. This model is used to impute the missing values. In our case we used an implicit model that is based on the hot-deck method. This is described further below.

7.5 Estimation Task

We define the posterior distribution of q as $\Pr\left(q \mid X, Y_{obs}\right)$.[21] However, the only function of q that is needed for the imputation task is $q_{Y|X}$. Therefore, during the estimation task, we draw repeated values of $q_{Y|X}$ from its posterior distribution $\Pr\left(q_{Y|X} \mid X, Y_{obs}\right)$. Let's call a drawn value $q^{*}_{Y|X}$.

7.6 Imputation Task

The posterior predictive distribution of the missing data given the observed data is defined by the following result:

$$\Pr\left(Y_{mis} \mid X, Y_{obs}\right) = \int \Pr\left(Y_{mis} \mid X, Y_{obs}, q\right) \Pr\left(q \mid X, Y_{obs}\right) dq \qquad \textbf{Eqn. 10}$$

We therefore draw a value of Y_{mis} from its conditional posterior distribution given $q^{*}_{Y|X}$. For example, we can draw $q^{*}_{Y|X} = (b^{*}, s^{*2})$ and compute the missing y_i from $f\left(y_i \middle| x_i, q^{*}_{Y|X}\right)$. This is the value that is imputed. This process is repeated M times.

7.7 Analysis Task

For each of the M complete data sets, we can calculate the value of Q. This provides us with the complete-data posterior distribution of Q: $\Pr\left(Q \mid X, Y_{obs}, Y_{mis}\right)$.

7.8 Combination Task

The basic result provided by Rubin [70] is:

$$\Pr\left(Q \mid X, Y_{obs}\right) = \int \Pr\left(Q \mid X, Y_{obs}, Y_{mis}\right) \Pr\left(Y_{mis} \mid X, Y_{obs}\right) dY_{mis} \qquad \textbf{Eqn. 11}$$

This result states that the actual posterior distribution of Q is equal to the average over the repeated imputations. Based on this result, a number of inferential procedures are defined.

The repeated imputation estimate of Q is:

$$\overline{Q} = \sum \frac{\hat{Q}_m}{M} \qquad \textbf{Eqn. 12}$$

which is the mean value across the M analyses that are performed.

The variability associated with this estimate has two components. First there is the within-imputation variance:

$$\overline{U} = \sum \frac{U_m}{M} \qquad \textbf{Eqn. 13}$$

and second the between imputation variance:

$$B = \frac{\sum \left(\hat{Q}_m - \overline{Q}\right)^2}{M - 1} \qquad \textbf{Eqn. 14}$$

[21] We use the notation $\Pr(\cdot)$ to denote a probability density.

The total variability associated with $\overline{Q}$ is therefore:

$$T = \overline{U} + \left(1 + M^{-1}\right)B \qquad \textbf{Eqn. 15}$$

In the case where Q is scaler, the following approximation can be made:

$$\frac{\left(Q - \overline{Q}\right)}{\sqrt{T}} \sim t_v \qquad \textbf{Eqn. 16}$$

where t_v is a t distribution with v degrees of freedom where:

$$v = \left(M - 1\right)\left(1 + r^{-1}\right)^2 \qquad \textbf{Eqn. 17}$$

and

$$r = \frac{\left(1 + M^{-1}\right)B}{\overline{U}} \qquad \textbf{Eqn. 18}$$

If one wants to test the null hypothesis that $H_0 : Q = 0$ then the following value can be referred to a t distribution with v degrees of freedom:

$$\frac{\overline{Q}}{\sqrt{T}} \qquad \textbf{Eqn. 19}$$

7.9 Hot-Deck Imputation: Overview

We will first start by presenting the hot-deck imputation procedure in general, then show the particular form of the procedure that we use in our analysis, and how this is incorporated into the multiple imputation process presented above.

Hot-deck procedures are used to impute missing values. They are a duplication approach whereby a *recipient* with a missing value receives a value from a *donor* with an observed value [29]. Therefore the donor's value is duplicated for each recipient. As can be imagined, this procedure can be operationalised in a number of different ways.

A basic approach for operationalising this is to sample from the n_{obs} observed values and use these to impute the n_{mis} missing values [58], where $n = n_{mis} + n_{obs}$. A simple sampling scheme could follow a multinomial model with sample size n_{mis} and probabilities $\left(1/n_{obs}, \ldots, 1/n_{obs}\right)$. It is more common, however, to use the X covariates to perform a poststratification. In such a case, the covariates are used to construct C disjoint classes of observations such that the observations within each class are as homogeneous as possible. This also has the advantage of further reducing nonresponse bias.

For example, if X consists of two binary vectors, then we have 4 possible disjoint classes. Within each class there will be some observations with Y observed and some with Y missing. For each of the missing values, we can randomly select an observed Y value and use it for imputation. This may result in the same observation serving as a donor more than once [77]. Here it is assumed that within each class the respondents follow the same distribution as the nonrespondents.

7.10 Metric-Matching Hot-Deck

It is not necessary that the X covariates are categorical. They can be continuous or a mixture of continuous and categorical variables. In such a case a distance function is defined, and the l nearest observations with the Y value observed serve as the donor pool [77].

An allied area where such metric-matching has received attention is the construction of matched samples in observational studies [67]. This is particularly relevant to our case because we cannot ensure in general that all the covariates that will be used in all analyses will be categorical. For the sake of brevity, we will only focus on the particular metric-matching technique that we employ.

7.11 Response Propensity Matching

In many observational studies[22] (see [12]) a relatively small group of subjects is exposed to a treatment, and there exists a larger group of unexposed subjects. Matching is then performed to identify unexposed subjects who serve as a control group. This is done to ensure that the treatment and control groups are both similar on background variables measured on all subjects.

Let the variable R_i denote whether a subject i was exposed ($R_i = 1$) or unexposed ($R_i = 0$) to the treatment. Define the propensity score, $e(X)$ as the conditional probability of exposure given the covariates (i.e., $e(X) = \Pr(R = 1 \mid X)$). Rosenbaum and Rubin [66] prove some properties of the propensity score that are relevant for us.

First, they show that the distribution of X is the same for all exposed and unexposed subjects within strata with constant values of $e(X)$. Exact matching will therefore tend to balance the X distributions for both groups. Furthermore, they also show that the distribution of the outcome variable Y is the same for exposed and unexposed subjects with the same value of $e(X)$ (or within strata of constant $e(X)$).

David et al. [15] adopt these results to the context of dealing with nonresponse in surveys. We can extrapolate and let $R_i = 1$ indicate that there was a response on Y for observation i, and that $R_i = 0$ indicates nonresponse. Hence we are dealing with response propensity as opposed to exposure propensity. We shall denote response propensity with $p(X)$. It then follows that under ignorable nonresponse if we can define strata with constant $p(X)$ then the distribution of X and Y are the same for both respondents and nonrespondents within each stratum.

To operationalise this, we need to address two issues. First, we need to estimate $p(X)$. Second, it is unlikely that we would be able to define sufficiently large strata where $p(X)$ is constant, and therefore we need to approximate this.

If we take the response indicator R to be Bernoulli random variable independently distributed across observations, then we can define a logistic regression model [39]:

$$p(X) = \frac{e^{(a_0 + a_1 X_1 + \cdots + a_{q-1} X_{q-1})}}{1 + e^{(a_0 + a_1 X_1 + \cdots + a_{q-1} X_{q-1})}}$$

This will provide us with an estimate of response propensity for respondents and nonrespondents.

We can then group the estimated response propensity into C intervals, with bounding values $0, p_1, p_2, \ldots, p_{C-1}, 1$. Strata can then be formed with observation i in stratum c if $p_{c-1} < p_i < p_c$ with $c = 1 \ldots C$. Therefore, we have constructed strata with approximately constant values of response

[22] These are studies where there is not a random assignment of subjects to treatments. For example, in the case of studying the relationship between exposure to cigarette smoke and cancer, it is not possible to deliberately expose some subjects to smoke.

propensity. In our application we set $C = 5$, dividing the estimated response propensity score using quintiles.

7.12 An Improper Hot-Deck Imputation Method

Now that we have constructed homogeneous strata, we can operationalise the metric-matching hot-deck imputation procedure by sampling with equal probability from the respondents within each stratum, and use the drawn values to impute the nonrespondent values in the same stratum. However, doing so we do not draw q from its posterior distribution, and then draw Y_{mis} from its posterior conditional distribution given the drawn value of q. Such a procedure would be improper and therefore some alternatives are considered, namely the approximate Bayesian bootstrap.

7.13 The Approximate Bayesian Bootstrap

A proper imputation approach that has been proposed is the Approximate Bayesian Bootstrap – ABB – (see [72][73]). This is an approximation of the Bayesian Bootstrap [69] that is easier to implement. The procedure for the ABB is, for each stratum, to draw with replacement z_{obs} Y values, where z_{obs} is the number of observed Y values in the stratum. Then, draw from that z_{mis} Y values with replacement, where z_{mis} is the number of observations with missing values in the stratum. The latter draws are then used to impute the missing values within the stratum. The drawing of z_{mis} missing values from a possible sample of z_{obs} values rather than from the actual observed values generates the appropriate between-imputation variability. This is repeated M times to generate multiple imputations.

7.14 Summary

The procedure that we have described implements multiple-imputation through the hot-deck method. It consists of constructing a response propensity model followed by an Approximate Bayesian Bootstrap.

This procedure is general and can be applied to impute missing values that are continuous or categorical. We have described it here in the context of univariate Y, but it is generally applicable to multivariate Y (see [70] for a detailed discussion of multiple-imputation for multivariate Y).

8 Acknowledgements

We wish to thank Thomas Kiesgen for helping with constructing the mappings from the C4 TRG document. We also wish to think all the participants in the SPICE Trials, without whom this study would not have been possible. In particular, our gratitude is given to Inigo Garro, Angela Tuffley, Bruce Hodgen, Alastair Walker, Stuart Hope, Robin Hunter, Dennis Goldenson, Terry Rout, Steve Masters, Ogawa Kiyoshi, Victoria Hailey, Peter Krauth, and Bob Smith.

9 References

[1] B. Baker, C. Hardyck, and L. Petrinovich: “Weak Measurements vs. Strong Statistics: An Empirical Critique of S. S. Stevens' Proscriptions on Statistics”. In *Educational and Psychological Measurement*, 26:291-309, 1966.

[2] S. Benno and D. Frailey: “Software Process Improvement in DSEG: 1989-1995”. In *Texas Instruments Technical Journal*, 12(2):20-28, March-April 1995.

[3] A. Bicego, M. Khurana, and P. Kuvaja: “Bootstrap 3.0: Software Process Assessment Methodology”. In *Proceedings of SQM’98*, 1998.

[4] G. Bohrnstedt and T. Carter: “Robustness in Regression Analysis”. In *Sociological Methodology*, H. Costner (ed.), Jossey-Bass, 1971.

[5] L. Briand, K. El Emam, and S. Morasca: "On the Application of Measurement Theory in Software Engineering". In *Empirical Software Engineering: An International Journal*, 1(1):61-88, Kluwer Academic Publishers, 1996.

[6] J. Brodman and D. Johnson: "What Small Businesses and Small Organizations Say about the CMM". In *Proceedings of the 16th International Conference on Software Engineering*, pages 331-340, 1994.

[7] J. Brodman and D. Johnson: "Return on Investment (ROI) from Software Process Improvement as Measured by US Industry". In *Software Process: Improvement and Practice*, Pilot Issue, John Wiley & Sons, 1995.

[8] K. Butler: "The Economic Benefits of Software Process Improvement". In *Crosstalk*, 8(7):14-17, July 1995.

[9] D. Card: "Understanding Process Improvement". In *IEEE Software*, pages 102-103, July 1991.

[10] B. Clark: *The Effects of Software Process Maturity on Software Development Effort*. PhD Thesis, University of Southern California, April 1997.

[11] F. Coallier, J. Mayrand, and B. Lague: "Risk Management in Software Product Procurement". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.

[12] W. Cochran. *Planning and Analysis of Observational Studies*. John Wiley & Sons, 1983.

[13] J. Cohen: *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[14] L. Cronbach: "Coefficient Alpha and the Internal Structure of Tests". In *Psychometrika*, September, pages 297-334, 1951.

[15] M. David, R. Little, M. Samuhel, and R. Triest: "Imputation Models Based on the Propensity to Respond". In *Proceedings of the Business and Economics Section*, American Statistical Association, pages 168-173, 1983.

[16] C. Deephouse, D. Goldenson, M. Kellner, and T. Mukhopadhyay: "The Effects of Software Processes on Meeting Targets and Quality". In *Proceedings of the Hawaiian International Conference on Systems Sciences*, vol. 4, pages 710-719, January 1995.

[17] R. Dion: "Elements of a Process Improvement program". In *IEEE Software*, 9(4):83-85, July 1992.

[18] R. Dion: "Process Improvement and the Corporate Balance Sheet". In *IEEE Software*, 10(4):28-35, July 1993.

[19] S. Dutta and L. van Wassenhove: "An Empirical Study of Adoption Levels of Software Management Practices within European Firms". INSEAD Research Initiative in Software Excellence Working paper, 1997.

[20] K. El Emam: "The Internal Consistency of the ISO/IEC 15504 Software Process Capability Scale". In *Proceedings of the 5th International Symposium on Software Metrics*, pages 72-81, IEEE CS Press, 1998.

[21] K. El Emam and D. R. Goldenson: "SPICE: An Empiricist's Perspective". In *Proceedings of the Second IEEE International Software Engineering Standards Symposium*, pages 84-97, August 1995.

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

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

[24] K. El Emam, J-N Drouin, and W. Melo (eds.): *SPICE: The Theory and Practice of Software Process Improvement and Capability Determination*. IEEE CS Press, 1998.

[25] K. El Emam, J-M Simon, S. Rousseau, and E. Jacquet: "Cost Implications of Interrater Agreement for Software Process Asssssments". In *Proceedings of the 5th International Symposium on Software Metrics*, pages 38-51, IEEE CS Press, 1998.

[26] K. El Emam and L. Briand: "Costs and Benefits of Software Process Improvement". In R. Messnarz and C. Tully (eds.): *Better Software Practice for Business Benefit: Principles and Experience*. IEEE CS Press, 1999.

[27] K. El Emam and D. Goldenson: "An Empirical Review of Software Process Assessments". In *Advances in Computers*, (to appear) 2000.

[28] R. Flowe and J. Thordahl: *A Correlational Study of the SEI's Capability Maturity Model and Software Development Performance in DoD Contracts*. MSc Thesis, The Air Force Institute of Technology, 1994.
[29] B. Ford: "An Overview of Hot-Deck Procedures". In W. Madow, I. Olkin, and D. Rubin (eds.): *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*. Academic Press, 1983.
[30] C. Franz and D. Robey: "Organizational Context, User Involvement, and the Usefulness of Information Systems". In *Decision Sciences*, 17:329-356, 1986.
[31] P. Fusaro, K. El Emam, and B. Smith: "The Internal Consistencies of the 1987 SEI Maturity Questionnaire and the SPICE Capability Dimension". In *Empirical Software Engineering: An International Journal*, 3:179-201, Kluwer Academic Publishers, 1997.
[32] D. Galletta and A. Lederer: "Some Cautions on the Measurement of User Information Satisfaction". In *Decision Sciences*, 20:419-438, 1989.
[33] P. Gardner: "Scales and Statistics". In *Review of Educational Research*, 45(1):43-57, Winter 1975.
[34] D. R. Goldenson and J. D. Herbsleb: "After the Appraisal: A Systematic Survey of Process Improvement, its Benefits, and Factors that Influence Success". Technical Report, CMU/SEI-95-TR-009, Software Engineering Institute, 1995.
[35] D. Goldenson, K. El Emam, J. Herbsleb, and C. Deephouse: "Empirical Studies of Software Process Assessment Methods". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*. IEEE CS Press, 1999.
[36] A. Gopal, T. Mukhopadhyay and M. Krishnan: "The Role of Software Processes and Communication in Offshore Software Development". Submitted for Publication, 1997.
[37] W. Harmon: "Benchmarking: The Starting Point for Process Improvement". In *Proceedings of the ESI Workshop on Benchmarking and Software Process Improvement*, April 1998.
[38] J. Herbsleb, A. Carleton, J. Rozum, J. Siegel, and D. Zubrow: "Benefits of CMM-Based Software Process Improvement: Initial Results". Technical Report, CMU-SEI-94-TR-13, Software Engineering Institute, 1994.
[39] D. Hosmer and S. Lemeshow: *Applied Logistic Regression*. John Wiley and Sons, 1989.
[40] W. Humphrey: "Characterizing the Software Process: A Maturity Framework". In *IEEE Software*, pages 73-79, March 1988.
[41] W. Humphrey, T. Snyder, and R. Willis: "Software Process Improvement at Hughes Aircraft". In *IEEE Software*, pages 11-23, July 1991.
[42] M. Ibanez and H. Rempp: "European User Survey Analysis". ESPITI project report, February 1996. (available from <http://www.esi.es>).
[43] B. Ives, M. Olson, and J. Baroudi: "The Measurement of User Information Satisfaction". In *Communications of the ACM*, 26(10):785-793, 1983.
[44] C. Jones: *Assessment and Control of Software Risks*. Prentice-Hall, 1994.
[45] C. Jones: "The Pragmatics of Software Process Improvements". In *Software Process Newsletter*, IEEE Computer Society TCSE, No. 5, pages 1-4, Winter 1996. . (available at http://www.seg.iit.nrc.ca/SPN)
[46] C. Jones: "The Economics of Software Process Improvements". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.
[47] F. Kerlinger: *Foundations of Behavioral Research*, Holt, Rinehart, and Winston, 1986.
[48] E. Kim and J. Lee: "An Exploratory Contingency Model of User Participation and MIS Use". In *Information and Management*, 11:87-97, 1986.
[49] H. Krasner: "The Payoff for Software Process Improvement: What it is and How to Get it". In K. El Emam and N. H. Madhavji (eds.): *Elements of Software Process Assessment and Improvement*, IEEE CS Press, 1999.
[50] M. Krishnan and M. Kellner: "Measuring Process Consistency: Implications for Reducing Software Defects". March 1998. Submitted for Publication.
[51] S. Labovitz: "Some Observations on Measurement and Statistics". In *Social Forces*, 46(2):151-160, December 1967.
[52] S. Labovitz: "The Assignment of Numbers to Rank Order Categories". In *American Sociological Review*, 35:515-524, 1970.
[53] P. Lawlis, R. Flowe, and J. Thordahl: "A Correlational Study of the CMM and Software Development Performance". In *Software Process Newsletter*, IEEE TCSE, No. 7, pages 1-5, Fall 1996. (available at http://www.seg.iit.nrc.ca/SPN)

[54] L. Lebsanft: "Bootstrap: Experiences with Europe's Software Process Assessment and Improvement Method". In *Software Process Newsletter*, IEEE Computer Society, No. 5, pages 6-10, Winter 1996. (available at http://www.seg.iit.nrc.ca/SPN)
[55] J. Lee and S. Kim: "The Relationship between Procedural Formalization in MIS Development and MIS Success". In *Information and Management*, 22:89-111, 1992.
[56] R. Lindsay and A. Ehrenberg: "The Design of Replicated Studies". In *The American Statistician*, 47(3):217-228, 1993.
[57] W. Lipke and K. Butler: "Software Process Improvement: A Success Story". In *Crosstalk*, 5(9):29-39, September 1992.
[58] R. Little and R. Rubin: *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987.
[59] F. McGarry, S. Burke, and B. Decker: "Measuring the Impacts Individual Process Maturity Attributes Have on Software Projects". In *Proceedings of the 5th International Software Metrics Symposium*, pages 52-60, 1998.
[60] J. McIver and E. Carmines: *Unidimensional Scaling*, Sage Publications, 1981.
[61] J. McKeen, T. Guimaraes, and J. Wetherbe: "The Relationship Between User Participation and User Satisfaction: An Investigation of Four Contingency Factors". In *MIS Quarterly*, pages 427-451, December 1994.
[62] J. Nunnally and I. Bernstein: *Psychometric Theory*. McGraw Hill, 1994.
[63] M. Paulk, B. Curtis, M-B Chrissis, and C. Weber: "Capability Maturity Model, Version 1.1". In *IEEE Software*, pages 18-27, July 1993.
[64] M. Paulk and M. Konrad: "Measuring Process Capability versus Organizational Process Maturity". In *Proceedings of the 4th International Conference on Software Quality*, October 1994.
[65] J. Rice: *Mathematical Statistics and Data Analysis*. Duxbury Press, 1995.
[66] P. Rosenbaum and D. Rubin: "The Central Role of the Propensity Score in Observational Studies for Causal Effects". In *Biometrika*, 70(1):41-55, 1983.
[67] P. Rosenbaum and D. Rubin: "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score". In *The American Statistician*, 39(1):33-38, 1985.
[68] R. Rosenthal: "Replication in Behavioral Research". In J. Neuliep (ed.): *Replication Research in the Social Sciences*. Sage Publications, 1991.
[69] D. Rubin: "The Bayesian Bootstrap". In *The Annals of Statistics*, 9(1):130-134, 1981.
[70] D. Rubin: *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.
[71] D. Rubin: "An Overview of Multiple Imputation". In *Proceedings of the Survey Research Section*, American Statistical Association, pages 79-84, 1988.
[72] D. Rubin and N. Schenker: "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse". In *Journal of the American Statistical Association*, 81(394):366-374, 1986.
[73] D. Rubin and N. Schenker: "Multiple Imputation in Health Care Databases: An Overview". In *Statistics in Medicine*, 10:585-598, 1991.
[74] H. Rubin: "Software Process Maturity: Measuring its Impact on Productivity and Quality". In *Proceedings of the International Conference on Software Engineering*, pages 468-476, 1993.
[75] H. Rubin: "Findings of the 1997 Worldwide Benchmark Project: Worldwide Software Engineering Performance Summary". Meta Group, 1998.
[76] D. Rubin, H. Stern, and V. Vehovar: "Handling 'Don't Know' Survey Responses: The Case of the Slovenian Plebiscite". In *Journal of the American Statistical Association*, 90(431):822-828. 1995.
[77] I. Sande: "Hot-Deck Imputation Procedures". In W. Madow and I. Olkin (eds.): *Incomplete Data in Sample Surveys, Volume 3: Proceedings of the Symposium*. Academic Press, 1983.
[78] J. Schaefer: *Analysis of Incomplete Multivariate Data*. Chapman & Hall, 1997.
[79] V. Sethi and W. King: "Construct Measurement in Information Systems Research: An Illustration in Strategic Systems". In *Decision Sciences*, 22:455-472, 1991.
[80] S. Siegel and J. Castellan: *Nonparametric Statistics for the Behavioral Sciences*, McGraw Hill, 1988.
[81] Software Engineering Institute: *The Capability Maturity Model: Guidelines for Improving the Software Process*. Addison Wesley, 1995.
[82] Software Engineering Institute: *C4 Software Technology Reference Guide – A Prototype*. Handbook CMU/SEI-97-HB-001, Software Engineering Institute, 1997.

[83] Software Engineering Institute: “Top-Level Standards Map”. Available at http://www.sei.cmu.edu/activities/cmm/cmm.articles.html, 23rd February 1998.
[84] Software Engineering Institute: “CMMI A Specification Version 1.1”. Available at http://www.sei.cmu.edu/activities/cmm/cmmi/specs/aspec1.1.html, 23rd April 1998.
[85] I. Sommerville and P. Sawyer: *Requirements Engineering: A Good Practice Guide*. John Wiley & Sons, 1997.
[86] P. Spector: “Ratings of Equal and Unequal Response Choice Intervals”. In *Journal of Social Psychology*, 112:115-119, 1980.
[87] The SPIRE Project: *The SPIRE Handbook: Better Faster Cheaper Software Development in Small Companies*. ESSI Project 23873, November 1998.
[88] S. Stevens: “Mathematics, Measurement, and Psychophysics”. In *Handbook of Experimental Psychology*, S. Stevens (ed.), John Wiley, 1951.
[89] A. Subramanian S. and Nilakanta: “Measurement: A Blueprint for Theory-Building in MIS”. In *Information and Management*, 26:13-20, 1994.
[90] D. Treiman, W. Bielby, and M. Cheng: “Evaluating a Multiple Imputation Method for Recalibrating 1970 U.S. Census Detailed Industry Codes to the 1980 Standard”. In *Sociological Methodology*, vol. 18, 1988.
[91] P. Velleman and L. Wilkinson: “Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading”. In *The American Statistician*, 47(1):65-72, February 1993.
[92] H. Wohlwend and S. Rosenbaum: “Software Improvements in an International Company”. In *Proceedings of the International Conference on Software Engineering*, pages 212-220, 1993.