Cost Implications of Interrater Agreement for Software Process Assessments

Khaled El Emam

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

Jean-Martin Simon

A.Q.T.
19, place de la Ferrandière
69003 Lyon
France
jms.aqt@wanadoo.fr

Sonia Rousseau [(1)]
Eric Jacquet[(2)]
(1)SANOFI Recherche
sonia.rousseau@tls1.elfsanofi.fr
(2)SANOFI Pharma
eric.jacquet@tls1.elfsanofi.fr
9, rue du Président S. Allende
94256 GENTILLY
France

International Software Engineering Research Network technical report ISERN-98-14

Cost Implications of Interrater Agreement for Software Process Assessments

Khaled El Emam

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

Jean-Martin Simon

A.Q.T.
19, place de la Ferrandière
69003 Lyon
France
jms.aqt@wanadoo.fr

Sonia Rousseau (1)
Eric Jacquet(2)
(1)SANOFI Recherche
sonia.rousseau@tls1.elfsanofi.fr
(2)SANOFI Pharma
eric.jacquet@tls1.elfsanofi.fr
9, rue du Président S. Allende
94256 GENTILLY
France

Abstract

Much empirical research has been done recently on evaluating and modeling interrater agreement in software process assessments. Interrater agreement is the extent to which assessors agree in their ratings of software process capabilities when presented with the same evidence and performing their ratings independently. This line of research was based on the premise that lack of interrater agreement can lead to erroneous decisions from process assessment scores. However, thus far we do not know the impact of interrater agreement on the cost of assessments. In this paper we report on a study that evaluates the relationship between interrater agreement and the cost of the consolidation activity in assessments. The study was conducted in the context of two assessments using the emerging international standard ISO/IEC 15504. Our results indicate that for organizational processes, the relationship is strong and in the expected direction. For project level processes no relationship was found. These results indicate that for assessments that include organizational processes in their scope, ensuring high interrater agreement could lead to a reduction in their costs.

1 Introduction

Software process assessments are an important and commonly used tool for focusing process improvement efforts and making improvement investments. A software process assessment is a subjective measurement procedure that involves expert judgement to identify quantitatively the process strengths and weaknesses of an organization. In addition, process assessments are intended to create a climate for change within an organization [5][6].

However, process assessments are believed to be an expensive tool. A recent article bemoans the high costs of process assessments [14]. An SEI analysis of feedback from CBA IPI[1] users found that more than a third of respondents who were assessment team leaders expressed concerns about the time it takes to conduct a CBA IPI, and so did 38% of the team members [4]. It therefore becomes an important research objective to identify ways of reducing the cost of assessments.

One approach for reducing the overall cost of assessments is to reduce the effort spent on the individual activities that make up an assessment. Our focus here is on reducing the cost of the consolidation activity. Commonly used team-based assessment methods have a consolidation activity[2] (e.g., see [5][2]). The consolidation activity is used to build consensus amongst the assessors on the evidence and findings of the assessment, as well as on the final ratings.

Respondents in the SEI study complained about insufficient time for consolidating evidence, reaching consensus, and/or crafting findings [4]. Therefore, there is a demand for reducing the costs of assessments in general, and specifically for reducing the cost of the consolidation activity. To address this, it is necessary to either increase the amount of time allocated for consolidation (by sacrificing other activities or increasing the total assessment cost), or reducing the amount of effort necessary to perform consolidation. Clearly the latter option is preferable.

---

[1] CMM Based Appriasal for Internal Process Improvement.

[2] This also is sometimes a series of activities during an assessment.

It is logical to expect that the more disagreement amongst the assessors, the more effort will be spent on consolidation. Therefore, reducing disagreement could lead to a reduction in consolidation effort[3].

Despite recent claims about the lack of evidence on interrater agreement [16], a number of studies have actually investigated the extent of agreement amongst assessor ratings during software process assessments[4] [9][11][12][15], and factors that influence agreement [10][19]. The basic premise behind this program of research has been that erroneous decisions can be made based on assessment scores that have low reliability [8].

However, we do not know if there is a relationship between agreement and the cost of assessments, and if so, what is the nature and magnitude of this relationship. If a relationship is found, then this provides an avenue for reducing the costs of assessments. Furthermore, such evidence would provide even more compelling reasons for ensuring high agreement amongst assessors.

The objective of this paper is to report on a study that investigates the economic impact of interrater agreement. In particular, we empirically evaluate the relationship between interrater agreement and the cost of the consolidation activity in software process assessments. The emerging ISO/IEC 15504 international standard on software process assessment was used during this study.

Briefly, our results indicate that there is a strong relationship between interrater agreement and consolidation effort for organizational-level processes. This is not the case for project-level processes. The implication of this result is that ensuring high interrater agreement can potentially reduce the cost of assessments if organizational-level processes are within the scope of the assessment. Therefore, studies such as [10][19] that investigate factors that influence agreement can potentially provide useful guidance for reducing the cost of assessments.

In the following section we provide a brief overview of the assessment model that was used in our study, as well as a specification of the cost model that we test. Section 3 is a description of our research method. In Section 4 we present the results, their interpretation, and their limitations are discussed. We conclude in Section 5 with a summary and directions for future work.

2 Background

The assessment model that was used in our study comes from the emerging international standard for software process assessment ISO/IEC 15504 [7]. The particular version that was used is the PDTR (Proposed Draft Technical Report). This version of the document set has recently undergone a series of empirical evaluations (described in [25]). The current study is part of this overall program of research. We first provide an overview of the relevant elements of ISO/IEC PDTR 15504, and then give a specification of the cost model that we test.

2.1 Overview of ISO/IEC PDTR 15504

The ISO/IEC 15504 architecture is two dimensional. Each dimension represents a different perspective on software process management. The first is the process dimension, and the second is the capability dimension. The process dimension is divided up into five process categories. Within each category is a set of processes. Each process is characterized by a process purpose. The processes that are within the scope of our study are presented later in this paper.

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. In our current study we only consider the first three levels, therefore we show the definition of the first three in Table 1. In Level 1, one attribute is directly rated. There are 2 attributes in each of the remaining 2 levels. The attributes are also shown in Table 1 (also see [7]).

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 2.

---

[3] Even perfect agreement is not expected to *eliminate* consolidation effort since some activities are performed during consolidation that would not be affected by agreement (such as summarizing the evidence). It is only that some of these activities will take longer the greater the disagreement.

[4] Specifically, these studies investigated interrater agreement. Interrater agreement is the extent to which assessors agree in their ratings of software process capabilities when presented with the same evidence and performing their ratings independently.

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 [7].

The scope of an assessment is an Organizational Unit (OU) [7]. 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 organization, although in a small organization the OU may be the whole organization. An OU may be, for example, a specific project or set of (related) projects, a unit within an organization focused on a specific life cycle phase (or phases), or a part of an organization responsible for all aspects of a particular product or product set.

ID	Definition
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 organization recognize 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 time scales. 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**

Table 1: Overview of the capability levels and attributes up to level 3.

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 2: The four-point attribute rating scale.

2.2 Model Specification

An assessment typically consists of a number of phases (for an overview see [27]). The total effort of an assessment is a sum of the effort consumed in each one of these phases. In assessments that include an explicit consolidation activity[5], the consolidation effort contributes towards the total assessment effort. Therefore, a reduction in the consolidation effort would be expected to lead to a reduction in the overall assessment effort.

During a team-based assessment, assessors are exposed to the same evidence. This evidence can be the result of pre-onsite questionnaires, the responses to questions during an interview, or from the inspection of documentation. Each assessor would maintain their own assessment record that documents the evidence that they observed. For example, members of an assessment team make their own notes during an interview and these become part of their individual assessment records. During consolidation, these notes are compared and integrated. If there is disagreement in the assessment records then some extra effort is spent on reaching consensus. This can include the re-examination of the assessment records, re-examination of evidence (e.g., re-inspect documents), and/or the collection of new evidence (e.g., interview people again).

Similarly, in some assessments, the assessors make their own preliminary process ratings[6] based on the interpretation of their assessment record. These are then discussed during the consolidation activity, and a consensus rating is made by the assessment team. If there is disagreement in the ratings, then it is expected that more effort will be spent reaching consensus.

Therefore, disagreements in the preliminary process ratings made by assessors may be due to one or a combination of reasons:

a) The assessment records of the assessors are different. This means that when presented with the same evidence assessors understand/capture it differently. When assessors make judgments or ratings based on inconsistent assessment records, then they are likely to disagree.

b) Assessors cannot distinguish between adjacent points on the rating scale (see Table 2). This means that even if the assessors have consistent assessment records, if they cannot distinguish between categories then they are more likely to disagree in their judgments or ratings.

c) Assessors may be biased. This means that even if their assessment records are consistent, they will make different judgments and ratings because they systematically interpret the scale categories differently.

All of the above may be caused by assessors' lack of experience, lack of knowledge of the assessment model, lack of knowledge of the capability scale, and/or their lack of knowledge of the OU and its business. They can also be caused by ambiguities in the assessment model and the capability measurement scale, as well as weaknesses in the assessment method itself.

[5] This is usually the case when the assessment is team based, for example, see [2][5].

[6] In some assessment method the assessors do not explicitly make independent ratings, but they do form opinions and do make judgments nevertheless.

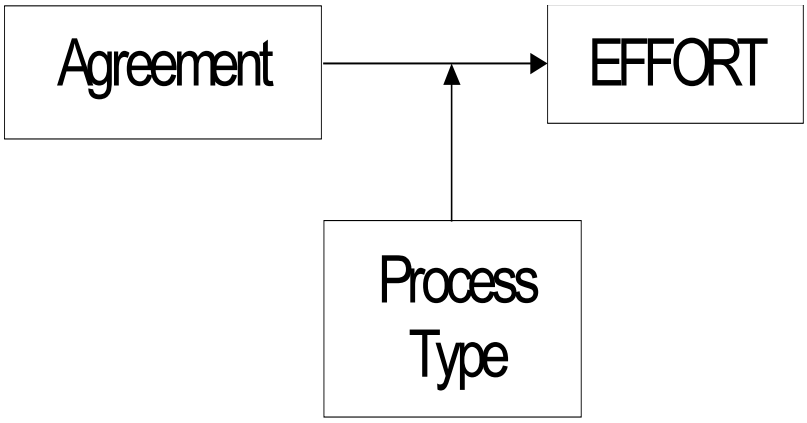

Figure 1: Path diagram showing the model being tested in the current study.

The basic hypothesis that we are testing is that there is a relationship between the extent of agreement in the ratings (or judgments) made by assessors independently and the cost of the consolidation phase. This means that we expect that a substantial proportion of variation in the cost of consolidation can be explained by the extent of interrater agreement.

The above relationship seems intuitively obvious. However, we also expect that this effect to be different for different types of processes. Consolidation may be easier for some processes than for others. Therefore, process type acts as a moderator variable on the relationship between agreement and consolidation cost.

Although other ways of identifying process types are plausible, we consider a process to be either an “Organization” or a “Project” process. Organizational processes are continuous activities that would span multiple projects. For example, the process “Provide Software Engineering Infrastructure” covers all projects with an OU, or “Engineer the Business” which includes providing an effective organizational culture. Project processes can be performed at a project level for every project (i.e., there could be an instance of the process for each project). For example, the process “Manage the project” has one instance per project. This distinction fits well with the manner in which many assessments using ISO/IEC PDTR 15504 are organized (i.e., by identifying individual projects to assess for “Project” type process, and by identifying individuals responsible for “Organizational” functions).

The final model that we test is depicted in Figure 1. This shows the relationship between Agreement and consolidation effort (EFFORT) moderated by Process Type.

3 Research Method

3.1 Description of Organization and Projects

In our study, we used data from two assessments that were conducted in France. In these assessments, the ISO/IEC PDTR 15504 documents were used. The company where the assessments was conducted is called Sanofi.

The Sanofi company belongs to the ELF Group. Its activities focus on drug research and production. All pharmaceutical molecules must undergo six to twelve long years of development from the moment of their discovery to the time they are given product license approval. Sanofi R&D has 2,500 employees, in nine units located in six countries (France, UK, Italy, Hungary, Spain and USA). From the research stage on the compound, to international commercialization, Sanofi R&D controls each phase to test scientifically both the indications for and the effects of the compounds.

The IS (Information Systems) department interacts with all of these activities as a support service. Computerized systems are necessary for several domains: discovery, preclinical studies, clinical investigation, and support. Development methods are either conventional (V model) or prototype based. Software packages are used extensively. The architecture is still "mainframe" for some systems, but mostly "Client-Server". The IS department manages the computerized systems life cycle from the initiation of a system to its retirement. They are used to working closely with users and with the support of the Research Quality Assurance.

	X1	X2	X3	Y1	Y2
Size of project in terms of effort	3 man-years	2.5 man-years	1 man-year	2 man-years	1 man-year
Programming language,	C, Visual Basic + off-the-shelf software	Third generation language	specific SQL	C, Visual Basic + off-the-shelf software	specific SQL
Development or maintenance projects	maintenance	maintenance	validation	maintenance	maintenance
Application domain	Electronic document management	data processing: collection, processing, visualization	data base , Client-server	Electronic document management	data base, Client-server

Table 3: Characteristics of assessed projects.

Process	Process Purpose	Number of Instances
Engineer the business	To provide the individuals in the organization and projects with a vision and culture that empowers them to function effectively	2
Define the process	To build a reusable library of process definitions (including standards, procedures, and models) that will support stable and repeatable performance of the software engineering and management process.	2
Improve the process	To improve continually the effectiveness and efficiency of the processes used by the organization in line with the business need, as a result of successful implementation of the process.	2
Provide skilled human resources	To provide the organization and projects with individuals who possess skills and knowledge to perform their roles effectively and to work together as a cohesive group.	2
Provide software engineering infrastructure	To provide a stable and reliable environment with an integrated set of software development methods and tools for use by the projects in the organization, consistent with and supportive of the defined process.	2
Supply software	To package, deliver, and install the software at the customer site and to ensure that quality software is delivered as defined by the requirements.	5
Operate software	To support the correct and efficient operation of the software for the duration of its intended usage in its installed environment.	4
Provide customer service	To establish and maintain an acceptable level of service to the customer to support effective use of the software	5
Maintain system and software	To manage modification, migration, and retirement of system components (such as hardware, software, manual operations, and network if any) in response to user requests.	4
Develop documentation	To develop and maintain documents, recording information produced by a process or activity within a process.	4
Perform configuration management	To establish and maintain the integrity of all the work products of a process or project.	4
Manage the project	To define the processes necessary to establish, coordinate, and manage a project and the resources necessary to produce a product.	4
Total		40

Table 4: Number of instances of each process assessed.

Two OUs within this company were assessed (called X and Y). A combination of organizational and project level processes were assessed in each OU. Three projects were assessed in the first OU and two projects in the second OU. The characteristics of these five projects are summarized in Table 3. The processes that were assessed and the number of instances in each are summarized in Table 4. The first five processes are classified as “Organization” type processes because they span multiple projects. The remaining processes are “Project” type processes, and instances are assessed for the projects in Table 3.

3.2 Description of Assessors

The same two assessors conducted both assessments. Both assessors met the minimal competence requirements stipulated in the ISO/IEC PDTR 15504 documents. Their experience and background is summarized in Table 5. Both assessors were external to the OU, and had conducted assessments together before.

	Assessor A	**Assessor B**
years in the software industry	14	3
years in process assessment and improvement	7	2
assessment methods & models they have experience with	ISO 9001, SPICE V1, and ISO/IEC PDTR 15504	ISO 9001, Bootstrap, and ISO/IEC PDTR 15504
number of SPICE-based assessments done in the past	6 (approximately 150 process instances)	3 (approximately 90 process instances)
internal vs. external to the organization	external	external

Table 5: Experience and background of assessors.

One recent study found that there are substantial differences in the ratings given by an experienced assessor (i.e., one who has conducted at least one assessment before) and an inexperienced assessor (i.e., one who has only received training on conducting an assessment, but did not conduct an actual assessment before) [19]. However, there was no substantial difference found between assessors who have conducted one or more assessments. In our study, both assessors have conducted at least one assessment before. Therefore we do not expect variation in agreement to be markedly influenced by the variation in assessment experience between our two assessors.

General Instructions for Conducting Interrater Agreement Studies

- For each process, divide the assessment team into two groups with at least one person per group.
- The two groups should be selected so that they both meet the minimal assessor competence requirements with respect to training, background, and experience.
- The two groups should use the same evidence (e.g., attend the same interviews, inspect the same documents, etc.), assessment method, and tools.
- The first group examining any physical artifacts should leave them as close as possible (organized/marked/sorted) to the state that the assessees delivered them.
- If evidence is judged to be insufficient, gather more evidence and both groups should inspect the new evidence before making ratings.
- The two groups independently rate the same process instances.
- After the independent ratings, the two groups then meet to reach consensus and harmonize their ratings for the final ratings profile.
- There should be no discussion between the two groups about rating judgment prior to the independent ratings.

Table 6: Guidelines for conducting interrater agreement studies.

3.3 Assessment Method

For collecting the necessary data, the assessment method follows the requirements for an interrater agreement study. These are: assessors must be exposed to the same evidence, their judgments must be made explicit in the form of ratings (so that we can calculate extent of agreement), and their ratings must be made independently of each other.

In this case, we divide the assessment team into two groups. In the current study, each of these groups had one assessor. Ideally both assessors should be equally competent in making attribute achievement ratings. In practice, both assessors need only meet minimal competence requirements since this is more congruent with the manner in which the 15504 documents would be applied. Each

assessor would be provided with the same information (e.g., all would be present in the same interviews[7] and provided with the same documentation to inspect), and then they would perform their ratings independently. Subsequent to the independent ratings, the two assessors would meet to reach a consensus or final assessment team rating. General guidelines for conducting interrater agreement studies are given in Table 6. The actual phases of the assessment method that was followed are summarized below.

3.3.1 Preparation Phase

As required by the ISO/IEC PDTR 15504, the assessment *input* is defined at the beginning of the assessment. This consists of, for example, the identity of the sponsor of the assessment and the sponsor's relationship to the OU being assessed; the assessment purpose including alignment with business goals; the assessment scope; the assessment constraints (e.g., availability of key resources, the maximum amount of time to be used for the assessment, specific processes or OUs to be excluded from the assessment); the identity of the assessors, including the competent assessor responsible for the assessment; the identity of assesses and support staff with specific responsibilities for the assessment; and any additional information to be collected during the assessment to support process improvement. During preparation, an important issue is to collect data on the context of the OU since the result of the assessment can only be interpreted within this particular context.

3.3.2 Data Collection Phase

To conduct the assessment, the interview technique was used, as well as document examination. For all of the processes within the scope of the assessment, only capability levels 1 to 3 were covered.

3.3.3 Ratings Phase

Each assessor collected his own assessment record during the interview. At the end of the day, each assessor took some time to review his own record and to make the process attributes ratings. Therefore, a specific meeting is dedicated to consolidate the assessment record and to establish a consensus between the two assessors when some divergence arises for one or several attribute ratings. This aspect is very important since one of the assessors may have missed or misunderstood some information. In the case that both assessors have missed some information, the sponsor (or the interviewee(s)) is contacted to obtain the missing information.

3.3.4 Debriefing

At the end of the assessment week (the number of days may depend on the number of assessed processes), the 2 assessors present to the interviewees the main results of the assessment. The objectives of this presentation include obtaining feedback from the interviewees about the results of the assessment.

During this meeting, the interviewees have the opportunity to "negotiate" the results by, for example, presenting further evidence. At this time, the results are only presented using a graphical approach.

3.3.5 Reporting

For the final assessment report, the results (weaknesses and strengths) were synthesized per process at the OU level. This global analysis is completed with the detailed analysis result for every assessed process for the considered project. This report is sent to the sponsor for approval.

3.4 Measurement

To test our hypothesized model we need to define three measures. In the context of ISO/IEC PDTR 15504, the unit of analysis is a process instance. Therefore, all measures have to be at that unit of analysis.

---

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

3.4.1 Measuring Cost (EFFORT)

The cost that we are interested in is that of the consolidation activity. This is measured as total effort of assessors during consolidation per process instance in person-minutes. During the assessment, the assessors kept records of how much effort was spent consolidating the findings and making harmonized ratings per process instance.

3.4.2 Measuring Agreement (AGREE)

The extent of agreement is measured by the proportion of attribute ratings on which the assessors agree. Therefore, it is a number that varies from zero to one. This simple measure is appropriate for this kind of study since we want a measure that reflects the extent of consensus building that is necessary during consolidation, and the more attribute ratings on which there is disagreement, the more consensus building is expected.

3.4.3 Measuring Process Type (TYPE)

Processes in the assessment were classified as either "Organization" or "Project", depending on whether particular projects were identified for making the process instance ratings.

3.5 Data Analysis

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

$$EFFORT = a + (b_1 \times TYPE) + (b_2 \times AGREE) + (b_3 \times TYPE \times AGREE)$$

Since TYPE can take only two values, it was treated as a dummy variable in the regression model and coded 0 for "Organization" and 1 for "Project". Aiken and West [1] recommend the dummy coding scheme, as opposed to, for example, effects coding, when there is only one categorical variable in the model since it facilitates a simpler interpretation of the results.

The analytical procedure that we followed is described in detail in [17]. This allows us to answer three questions: (1) "is there an interaction effect?", (b) "if so, what is the strength of the effect?", and (3) "if so, what is the nature of the effect?".

To determine whether there is an interaction effect we utilize a hierarchial F test for the mul;tiple regression equation without the interaction term and one with the interaction term. If the F ratio is statistically significant, then we can interpret that as meaning that there is an interaction effect. The significance of b_3 coefficient determined using a t-test gives the same answer.

To answer the second question we compare the R^2 value of the model without the interaction term with the model with the interaction term. This gives us the increase in explained variation due to adding the interaction term.

To determine the nature of the effect, we calculate the two straight line equations for "Organization" and "Project" processes, and determine whether each of the slopes is different from zero by using a t-test[8]. For the latter we use Bonferroni adjusted alpha levels (see [22]) since two t-tests are performed, as suggested in [17]. The two equations are as follows:

$$EFFORT_{TYPE=Organization} = a + (b_2 \times AGREE)$$

$$EFFORT_{TYPE=\mathrm{Pr}oject} = (a + b_1) + ((b_2 + b_3) \times AGREE)$$

The alpha level that we used for our statistical analysis was 0.1. As noted in [17], the dummy variable should not be centered. However, for the continuous AGREE variable, it is expected that it would have a strong relationship with the interaction term. This potentially introduces multicollinearity in our regression model. To address this, the AGREE variable is centered by subtracting the mean from each raw value [1]. The interaction term is constructed using the centered AGREE variable.

[8] The computation of the standard error in the t-ratio in this context is provided in [17].

4 Results

4.1 Interaction Model

The final model that we obtained is depicted graphically in Figure 2. The model parameters are shown in Table 7. The goodness of fit as measured using R^2 is 0.63 (with p<0.000). This is quite a good fit given that only two variables are considered. Furthermore, in comparison with previous software process assessment research that developed interaction models [10], our 63% explained variation is approximately double, further indicating a rather good fit by reference to precedent.

The model parameters when there is centering of the continuous variable can be interpreted as follows. The *a* parameter is the consolidation cost for Organizational processes at the mean AGREE value for the whole sample (which is 0.713). The $(a + b_1)$ value is the consolidation cost for Project processes at the mean AGREE value. In this case, the b_1 parameter represents the distance in consolidation cost between the two types of processes at the mean AGREE value. The b_2 parameter is the slope of the line for Project type processes, and the b_3 parameter is the difference in slope between the two types of processes.

To revisit the issue of multicollinearity once more, we use the Variance Inflation Factor (VIF) to determine the extent of multicollinearity. This is a commonly used approach, and involves computing $\frac{1}{1-R_k^2}$ for each independent variable, where R_k^2 is the R^2 value from regressing the independent variable k on all the remaining k-1 independent variables [3][20]. Values of VIF greater than 10 are usually taken as an indication of multicollinearity problems. After centering, the largest VIF value for our model was 4.

As seen from Table 7, the regression parameter for the interaction term is statistically significant, therefore indicating that there is an interaction effect. The size of this effect was computed to be 0.34. This means that the interaction term explains an additional 34% of the variation in consolidation cost compared with the main effects only model. This is a large amount of explained variation compared to previous software engineering studies of complex organizational phenomenon [10][13]. When we tested the slope of each of the two lines in Figure 2 to see if they were different from zero, we found that for the "Organization" line the slope was significantly different from zero (p<0.000). Therefore, for Organization type processes, as agreement increases, less effort is spent on consolidation activities. But this was not the case for the "Project" line. This indicates that disagreement for "Project" type processes has no impact on the cost of assessments.

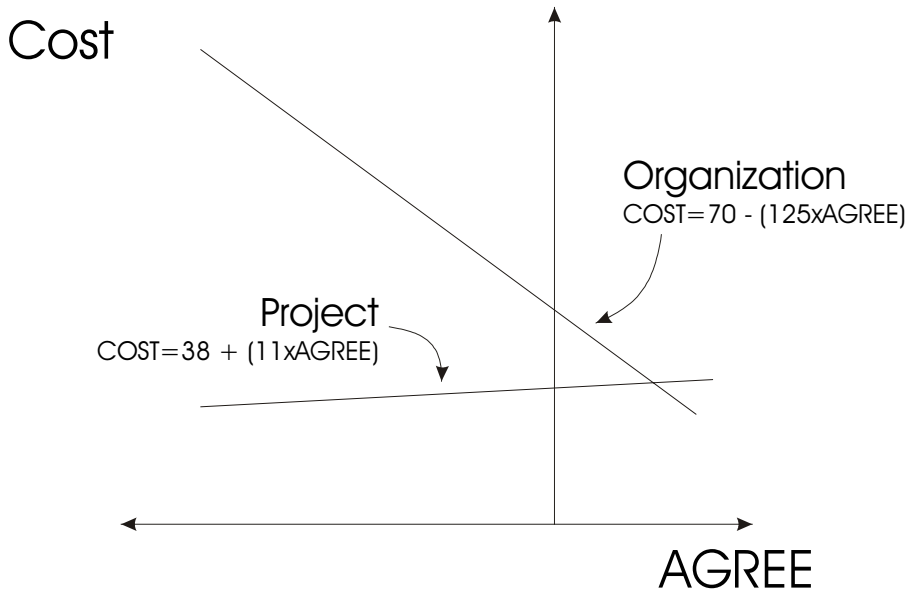

Figure 2: Graphical depiction of the consolidation cost model.

Regression Parameter	Value	p-value[9]
Intercept (a)	70	p<0.000
TYPE (b_1)	-32	p<0.000
AGREE (b_2)	-125	p<0.000
TYPE x AGREE (b_3)	136	p<0.000

Table 7: Parameters of the regression model with interaction.

For the range of the AGREE variable (0 to 1), the two lines would intersect at an AGREE value of approximately 0.95 (this value is 0.235 when AGREE is centered). We would like to identify the region around this intersection point where the differences between the two slopes are statistically significant. For any given value of AGREE we can do so using the Johnson-Neyman technique [18][23]. However, to determine *regions of significance*, we use Potthoff's extensions to the basic Johnson-Neyman technique [21]. The definition of regions of significance when applied to our context is that in the long run, not more than 10% of such regions will contain any point for which the two process types differ in expected consolidation effort. When we applied the procedure, the regions of significance were outside the possible range of AGREE. According to the guidance in [1], this indicates that the two regression lines differ for all values of AGREE since the b_1 parameter for the group effect was significant. However, the intersection occurs at an impossible value given our data set (AGREE of 0.95 is not plausible since it can only take values in 0.2 increments). Therefore, we can reasonably conclude that consolidation effort for organization processes tends to be higher except for the case when there is perfect agreement, at which point there does not seem to be a substantial difference between the two process types in terms of consolidation effort.

4.2 Threats to Validity

Our basic premise was that a reduction in consolidation cost will reduce the overall cost of an assessment. In our results we found a relationship between the extent of agreement and consolidation effort for "Organization" type processes. This would then suggest that increasing agreement will reduce the overall cost of an assessment that includes "Organization" processes in its scope. However, this conclusion is not necessarily justifiable unless we eliminate the possibility of *redistribution of effort* as an explanation of our findings. We must therefore also consider the "evidence collection" effort.

One can argue that during the assessment, if the assessors spend more effort collecting evidence, then they are more likely to agree in their judgments since there is more evidence to base a judgment upon. Then this high agreement will lead to less consolidation effort, at least for the "Organization" processes. In total then, there may be no impact on assessment cost because the low consolidation effort is *compensated* for by more evidence collection effort. If this argument is true then we would expect a strong positive relationship between evidence collection effort and agreement.

Conversely, one can argue that in reality there is a *reenforcing* relationship between evidence collection effort and agreement. It is plausible that assessors spend more effort collecting evidence on the process instances where they feel less confident in the repeatability of their judgments, which tend to be the ones they disagree on. Therefore, the extent of evidence collection effort is an indication of the extent of disagreement rather than being one of the causal factors that affects agreement. If this argument is true then we would expect the relationship between evidence collection effort and Agreement to be negative.

We evaluated the bivariate relationship between evidence collection effort[10] and Agreement. This was evaluated using the Pearson correlation coefficient [26]. We also used the parallel nonparametric Spearman's rho coefficient since it makes less assumptions, especially about the nature of the distributions of the two variables [24].

[9] All p-values are for two tailed t-tests on the regression parameter.

[10] During the assessments, evidence was collected per process instance (i.e., interviews and inspecting documents). The assessors kept records of the effort consumed doing so per process instance.

The results of the bivariate analysis are shown in Table 8. This indicates that, at a two-tailed alpha level of 0.1, the relationship is statistically significant and in the negative direction. We therefore interpret this to mean that the extent of effort spent on evidence collection is an indication of the extent of disagreement, rather than being a factor that leads to more/less agreement.

	Magnitude	**p-value**
Pearson Correlation	-0.289	p=0.074
Spearman Correlation[11]	-0.317	p=0.049

Table 8: Bivariate relationship between evidence collection effort and agreement.

This analysis eliminates the alternative interpretation of our results, and therefore suggests that increasing agreement would lead to a reduction in the overall cost of an assessment for "Organization" type processes.

Another issue of concern for this study is its external validity. External validity threats question the extent to which we can generalize our findings beyond the current study. In particular, one can argue that the results we obtained were specific to the two assessors that took part in our study using a particular assessment method. This threat can be addressed through replications of this study with other assessors and using different methods. We plan to do so within our program of research on the reliability of software process assessments.

4.3 Discussion

The results of our analysis suggest that it is certainly worthwhile to increase interrater agreement for "Organization" type processes. This can reduce the costs of assessments. Based on recent data collected about ISO/IEC PDTR 15504 based assessments conducted in Europe and Australia, 33% of assessments include organizational processes within their scope [25], suggesting that these findings are potentially applicable to one third of ISO/IEC 15504 assessments.

It is also clear from our results that the more effort spent on evidence collection does not lead to less consolidation effort since assessors tend to collect more evidence when they feel that they are not able to make repeatable judgments.

The implications are that assessors should pay most attention to the ratings of "Organization" type processes in order to reduce the cost of the assessment, and also that future research should focus on improving the reliability of rating this type of process.

5 Conclusions

In this paper we investigated the relationship between the interrater agreement of software process assessments and the cost of the consolidation phase. We found that for organizational level processes higher agreement is related to a reduction in cost, while for project level processes there is no relationship. This indicates that for process assessments that include organizational processes in their scope ensuring high agreement can lead to a reduction in their overall costs. Furthermore, given that previous studies highlight the resource constraint difficulties faced by assessors in performing consolidation activities, reducing the amount of effort required can alleviate some of these difficulties.

An interesting observation from this study was that the process type (i.e., organizational vs. project) can help explain variations in the cost of assessments. Future studies that construct empirical models of software process assessments in general, and specifically of cost, ought to consider this variable.

[11] For samples greater than 25, the value $\rho\sqrt{N-1}$ is approximately normally distributed with mean zero and standard deviation 1 and can be used for computing p-values. Alternatively a similar statistic that is approximately distributed as Student's t can be used [24]. In our case, the p value that is presented in the table is obtained from the Student t distribution. However, [24] also recommend using tabulated critical values which are computed using a permutation approach for sample sizes similar to ours. Our conclusions would not change if tabulated critical values are used.

6 Acknowledgements

This study is supported by ELF *Innovation* to promote software process assessment as a technique for Software Quality Management.

7 References

[1] L. Aiken and S. West: *Multiple Regression: Testing and Interpreting Interactions*. Sage Publications, 1991.

[2] R. Barbour: *Software Capability Evaluation Version 3.0: Implementation Guide for Supplier Selection*. Technical Report CMU/SEI-95-TR-012, Software Engineering Institute, 1996.

[3] S. Chatterjee and B. Price: *Regression Analysis by Example*. John Wiley & Sons, 1991.

[4] D. Dunaway, D. Goldenson, I. Monarch, and D. White: "How Well is CBA IPI Working? User Feedback". In *Proceedings of the SEPG Conference*, Chicago, 1998.

[5] D. Dunaway and S. Masters: *CMM-Based Appraisal for Internal Process Improvement (CBA IPI): Method Description*. Technical Report CMU/SEI-96-TR-7, Software Engineering Institute, 1996.

[6] K. Dymond: "Essence and Accidents in SEI-Style Assessments or 'Maybe This Time the Voice of the Engineer Will be Heard'". In *IEEE TCSE Software Process Newsletter*, 9:1-7, 1997.

[7] 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.

[8] 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, Canada, August 1995.

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

[10] K. El Emam, R. Smith, and P. Fusaro: "Modelling the Reliability of SPICE Based Assessments". In *Proceedings of the International Symposium on Software Engineering Standards*, pages 69-82, 1997.

[11] K. El Emam, L. Briand, and R. Smith: "Assessor Agreement in Rating SPICE Processes". In *Software Process Improvement and Practice Journal*, 2(4):291-306, John Wiley, 1997.

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

[13] 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..

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

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

[16] E. Gray and W. Smith: "On the Limitations of Software Process Assessment and the Recognitiion of a Required Re-Orientation for Global Process Improvement". In *Software Quality Journal*, 7:21-34, 1998.

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

[18] P. Johnson and L. Fay: "The Johnson-Neyman Technique, Its Theory and Application". In *Psychomterika*, 15(4):349-367, 1950.

[19] M. Khurana and K. El Emam: "Assessment Experience and the Reliability of Assessments". In *IEEE TCSE Software Process Newsletter*, No. 12, 1998.

[20] J. Neter, W. Wasserman, and M. Kutner: *Applied Statistical Models*. Irwin, 1990.

[21] R. Potthoff: "On the Johnson-Neyman Technique and Some Extensions Thereof". In *Psychometrika*, 29:241-256, 1964.

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

[23] D. Rogosa: "Comparing Nonparallel Regression Lines". In *Psychological Bulletin*, 88(2):307-321, 1980.

[24] S. Siegel and N. Castellan: *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[25] SPICE Project Trials Team: *Phase 2 Trials Interim Report*. March 24th, 1998.
[26] R. Wherry: *Contributions to Correlational Analysis*. Academic Press, 1984.
[27] S. Zahran: *Software Process Improvement: Practical Guidelines for Business Success*. Addison-Wesley, 1998.