SINTEF

SINTEF ICT

Address: NO-7465 Trondheim, NORWAY
Location: Forskningsveien 1
Telephone: +47 22 06 73 00
Fax: +47 22 06 73 50

Enterprise No.: NO 948 007 029 MVA

SINTEF REPORT

TITLE

Evaluation of Experiences from Applying the PREDIQT Method in an Industrial Case Study

AUTHOR(S)

Aida Omerovic, Bjørnar Solhaug and Ketil Stølen

CLIENT(S)

Research Council of Norway

REPORT NO.	CLASSIFICATION	CLIENTS REF.	
SINTEF A17562	Open	180052/S10	
CLASS. THIS PAGE	ISBN	PROJECT NO.	NO. OF PAGES/APPENDICES
Open	978-82-14-04970-1	90B245	21/7
ELECTRONIC FILE CODE		PROJECT MANAGER (NAME, SIGN.) Ketil Stølen	CHECKED BY (NAME, SIGN.) Gyrd Brændeland
FILE CODE	DATE	APPROVED BY (NAME, POSITION, SIGN.)	
	2011-01-03	Bjørn Skjellaug	

ABSTRACT

We have developed a method called PREDIQT for model-based prediction of impacts of architectural design changes on system quality. A recent case study indicated feasibility of the PREDIQT method when applied on a real-life industrial system. This paper reports on the experiences from applying the PREDIQT method in a second and more recent case study – on an industrial ICT system from another domain and with a number of different system characteristics, compared with the previous case study. The analysis is performed in a fully realistic setting. The system analyzed is a critical and complex expert system used for management and support of numerous working processes. The system is subject to frequent changes of varying type and extent. The objective of the case study has been to perform an additional and more structured evaluation of the PREDIQT method and assess its performance with respect to a set of success criteria. The evaluation argues that: 1) the PREDIQT-based analysis facilitates predictions providing sufficient understanding of the impacts of architectural design changes on system quality characteristics, so that informed decisions can be made; 2) the PREDIQT-based analysis is cost-effective; 3) the prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality; 4) the prediction models are sufficiently comprehensible to allow the domain experts to be actively involved in all phases of the PREDIQT process and achieve the goals of each phase with a common understanding of the results; and 5) the PREDIQT-based analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality. Moreover, the study has provided useful insights into the weaknesses of the method and suggested directions for future research and improvements.

Key words: Quality prediction, System architectural design, Change impact analysis, Modeling, Simulation

KEYWORDS	ENGLISH	NORWEGIAN
GROUP 1	Quality prediction, System architectural design, Empirical evaluation	Kvalitetsprediksjon, Systemarkitektur, Empirisk evaluering
GROUP 2	Modeling, Change impact analysis	Modellering, Konsekvensanalyse
SELECTED BY AUTHOR	Quality prediction, Architectural design	Kvalitetsprediksjon, Systemarkitektur

CONTENTS

Evaluation of Experiences from Applying the PREDIQT Method in an Industrial Case Study

Aida Omerovic[1,2], Bjørnar Solhaug[1] and Ketil Stølen[1,2]
[1]SINTEF ICT, P.O. Box 124, 0314 Oslo, Norway
[2]University of Oslo, Department of Informatics, P.O. Box 1080, 0316 Oslo, Norway
Email: {aida.omerovic, bjornar.solhaug, ketil.stolen}@sintef.no

***Abstract*—We have developed a method called PREDIQT for model-based prediction of impacts of architectural design changes on system quality. A recent case study indicated feasibility of the PREDIQT method when applied on a real-life industrial system. This paper reports on the experiences from applying the PREDIQT method in a second and more recent case study – on an industrial ICT system from another domain and with a number of different system characteristics, compared with the previous case study. The analysis is performed in a fully realistic setting. The system analyzed is a critical and complex expert system used for management and support of numerous working processes. The system is subject to frequent changes of varying type and extent. The objective of the case study has been to perform an additional and more structured evaluation of the PREDIQT method and assess its performance with respect to a set of success criteria. The evaluation argues that: 1) the PREDIQT-based analysis facilitates predictions providing sufficient understanding of the impacts of architectural design changes on system quality characteristics, so that informed decisions can be made; 2) the PREDIQT-based analysis is cost-effective; 3) the prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality; 4) the prediction models are sufficiently comprehensible to allow the domain experts to be actively involved in all phases of the PREDIQT process and achieve the goals of each phase with a common understanding of the results; and 5) the PREDIQT-based analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality. Moreover, the study has provided useful insights into the weaknesses of the method and suggested directions for future research and improvements.**

***Index Terms*—Quality prediction, System architectural design, Change impact analysis, Modeling, Simulation.**

I. Introduction

When adapting a system to new usage patterns, processes or technologies, it is necessary to foresee the implications that the architectural design changes have on system quality. Predictability with respect to non-functional requirements is one of the necessary conditions for the trustworthiness of a system. Examination of quality outcomes through implementation of the different architecture design alternatives is often unfeasible. A model-based approach is then an alternative. We have developed a method called PREDIQT with the aim to facilitate model-based prediction of impacts of architectural design changes on system quality. Examples of quality characteristics include availability, scalability, security and reliability.

A recent case study [16] indicated feasibility of the PREDIQT method when applied on a real-life industrial system. The promising empirical results and experiences from the previous case study encouraged further and more structured evaluation of the PREDIQT method. This paper addresses experiences from applying PREDIQT on another real-life industrial system from a different domain and with different system characteristics (lifetime, purpose, technology the system is implemented on, number of users and kind of users), compared to the previous case study.

The target system analyzed serves as a semantic model and a repository for representation of the system owner's core working processes and rules, and as a knowledge database. It is a business-critical and complex expert system used for management and support of numerous working processes, involving hundreds of professional users every day. The system is subject to frequent architectural design changes of varying type and extent. The system owner, who was also the client commissioning the analysis, required full confidentiality with respect to the kind of system targeted, the models obtained, the personnel involved and the name of the organization. This paper reports solely on the experiences obtained by the participants of the real-life case, describes the process undergone, the evaluation results, the observations and the properties of the artifacts. The reported experiences and results have provided valuable insight into the strengths and weaknesses of the method.

The case study was conducted in the year 2010. The first overall two phases of the PREDIQT method were conducted in their entirety, while the last phase was partially covered. In addition, the method is assessed through a thought experiment based evaluation of predictions and a postmortem review. All prediction models were developed during the analysis and the entire target system (within the predefined scope) was analyzed. The analysis was performed in the form of five workshops and six intermediate meetings in a fully realistic setting in terms of the scope, the objectives, the process, the prediction models and the participants.

The rest of the paper is structured as follows. We briefly present the PREDIQT method in Section II. The research method is summarized in Section III. Section IV presents five success criteria, which cover the needs of the three stakeholder groups, and which both the case study and the contents of this paper have primarily been driven by. The process undergone during the PREDIQT-based analysis is presented in Section V. Results of evaluation and a postmortem review are summarized

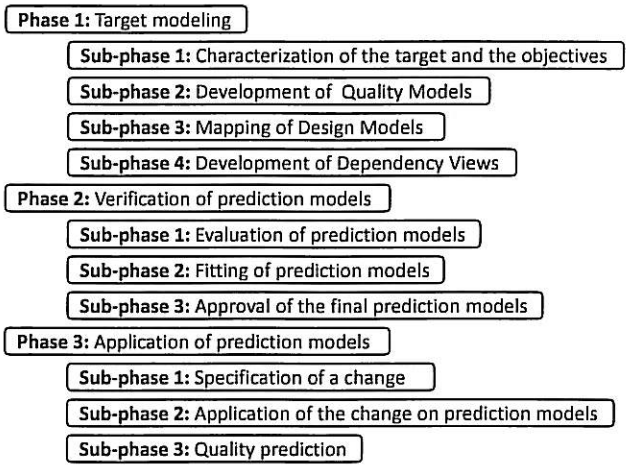

Fig. 1. A simplified overview of the process of the PREDIQT method

in Section VI. Section VII provides an evaluation of the experiences and results, with respect to the five pre-defined success criteria, before concluding in Section VIII. A more thorough presentation of the research method is provided in Appendix 1. Setup and data collection during the PREDIQT-based analysis are outlined in Appendix 2. The outcomes of the process, in terms of artifacts, evaluation results and observations, are reported in Appendix 3. Appendix 4 presents the design of the evaluation template used in relation to the postmortem review. A summary of the feedback received through the evaluation template is provided in Appendix 5. Threats to validity and reliability are discussed in Appendix 6. Appendix 7 summarizes some of the related work.

II. Overview of the PREDIQT method

The PREDIQT method produces and applies a multi-layer model structure, called prediction models, which represent system relevant quality concepts (through "Quality Models"), architectural design (through "Design Models"), and the dependencies between architectural design and quality (through "Dependency Views"). The Design Models are used to specify the target system and the changes whose effects on quality are to be predicted. The Quality Models are used to formalize the quality notions and define their interpretations. The values and the dependencies modeled through the DVs are based on the definitions provided by the Quality Models. The DVs express the interplay between the system architectural design and the quality characteristics. Once a change is specified on the Design Models, the affected parts of the DVs are identified, and the effects of the change on the quality values are automatically propagated at the appropriate parts of the DV. This section briefly outlines the PREDIQT method in terms of the process and the artifacts. For further details on PREDIQT, see [16, 17, 19].

The process of the PREDIQT method consists of three overall phases. Each phase is decomposed into sub-phases, as illustrated in a simplified form by Figure 1. Based on

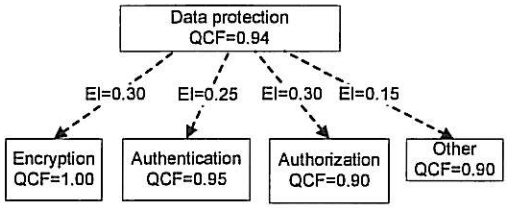

Fig. 2. Excerpt of an example DV with fictitious values

the initial input, the stakeholders involved deduce a high level characterization of the target system, its scope and the objectives of the prediction analysis, by formulating the system boundaries, system context (including the operational profile), system lifetime and the extent (nature and rate) of design changes expected. Quality Models are created in the form of a tree, by defining quality notions with respect to the target system. The Quality Models represent a taxonomy with interpretations and formal definitions of system quality notions. The total quality of the system is decomposed into characteristics, sub-characteristics and quality indicators. The Design Models represent the architectural design of the system.

For each quality characteristic defined in the Quality Model, a quality characteristic specific Dependency View (DV) is deduced from the Design Models and the Quality Models of the system under analysis. This is done by modeling the dependencies of the architectural design with respect to the quality characteristic that the DV is dedicated to, in the form of multiple weighted and directed trees. A DV comprises two notions of parameters:

1) EI: Estimated degree of Impact between two nodes, and
2) QCF: degree of Quality Characteristic Fulfillment.

Each arc pointing from the node being influenced is annotated by a quantitative value of EI, and each node is annotated by a quantitative value of QCF.

Figure 2 shows an excerpt of an example DV with fictitious values. In the case of the *Encryption* node of Figure 2, the QCF value expresses the goodness of encryption with respect to the quality characteristic in question, e.g., security. A quality characteristic is defined by the underlying system specific Quality Models, which may for example be based on the ISO 9126 product quality standard [1]. A QCF value on a DV expresses to what degree the node (representing system part, concern or similar) is realised so that it, within its own domain, fulfills the quality characteristic. The QCF value is based on the formal definition of the quality characteristic (for the system under analysis), provided by the Quality Models. The EI value on an arc expresses the degree of impact of a child node (which the arc is directed to) on the parent node, or to what degree the parent node depends on the child node, with respect to the quality characteristic under consideration.

"Initial" or "prior" estimation of a DV involves providing QCF values to all leaf nodes, and EI values to all arcs. Input to the DV parameters may come in different forms (e.g., from domain expert judgments, experience factories, measurements, monitoring, logs, etc.), during the different phases of

the PREDIQT method. The DV parameters are assigned by providing the estimates on the arcs and the leaf nodes, and propagating them according to the general DV propagation algorithm. Consider for example the *Data protection* node on Figure 2 (denoting: DP: Data protection, E: Encryption, AT: Authentication, AAT: Authorization, and O:Other):

$$QCF_{(DP)} = QCF_{(E)} \cdot EI_{(DP \rightarrow E)} + QCF_{(AT)} \cdot EI_{(DP \rightarrow AT)} + QCF_{(AAT)} \cdot EI_{(DP \rightarrow AAT)} + QCF_{(O)} \cdot EI_{(DP \rightarrow O)} \quad \text{Eq. 1}$$

The DV based approach constrains the QCF of each node to range between 0 and 1, representing minimal and maximal characteristic fulfillment (within the domain of what is represented by the node), respectively. This constraint is ensured through the formal definition of the quality characteristic rating (provided in the Quality Models). The sum of EIs, each between 0 (no impact) and 1 (maximum impact), assigned to the arcs pointing to the immediate children must be 1 (for model completeness purpose). Moreover, all nodes having a common parent have to be orthogonal (independent). The dependent nodes are placed at different levels when structuring the tree, thus ensuring that the needed relations are shown at the same time as the tree structure is preserved.

The general DV propagation algorithm, exemplified by Eq. 1, is legitimate since each quality characteristic DV is complete, the EIs are normalized and the nodes having a common parent are orthogonal due to the structure. A DV is complete if each node which is decomposed, has children nodes which are independent and which together fully represent the relevant impacts on the parent node, with respect to the quality characteristic that the DV is dedicated to.

The rationale for the orthogonality is that the resulting DV structure is tree-formed and easy for the domain experts to relate to. This significantly simplifies the parametrization and limits the number of estimates required, since the number of interactions between the nodes is minimized. Although the orthogonality requirement puts additional demands on the DV structuring, it has shown to represent a significant advantage during the estimation.

The "Verification of prediction models" phase aims to validate the prediction models (with respect to the structure and the individual parameters), before they are applied. A measurement plan with the necessary statistical power is developed, describing what should be evaluated, when and how. Both system-as-is and change effects should be covered by the measurement plan. Model fitting is conducted in order to adjust the DV structure and the parameters, to the evaluation results. The objective of the "Approval of the final prediction models" sub-phase is to evaluate the prediction models as a whole and validate that they are complete, correct and mutually consistent after the fitting. If the deviation between the model and the new measurements is above the acceptable threshold after the fitting, the target modeling is re-initiated.

The "Application of prediction models" presupposes that the prediction models are approved. During this phase, a specified change is applied to the Design Models and the DVs, and its effects on the quality characteristics at the various abstraction levels are simulated on the respective DVs. The change specification should clearly state all deployment relevant facts, necessary for applying the change. The "Apply the change on prediction models" phase involves applying the specified architectural design change on the prediction models. When an architectural design change is applied on the Design Models, it is according to the definitions in the Quality Model, reflected to the relevant parts of the DV. Thereafter, the DV provides propagation paths and quantitative predictions of the new quality characteristic values, by propagating the change throughout the rest of each one of the modified DVs, based on the general DV propagation algorithm. We have earlier developed tool support [16] based on MS Excel for simulation and sensitivity analysis related to the DVs.

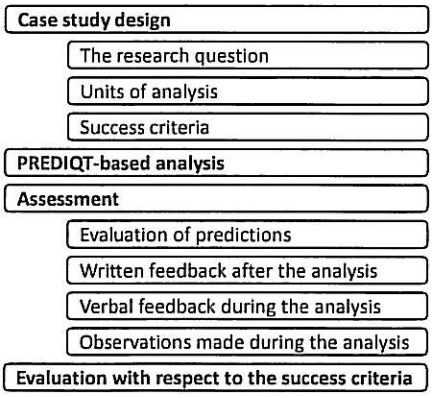

Fig. 3. Main stages of the research method

III. Research Method

The research method is motivated by the guidelines for case study research provided by Yin [21]. A deductive approach is undertaken, where the already defined PREDIQT method is exposed to an empirical trial in the form of a case study. For more details on the research method, see Appendix 1.

The main stages of the research method are depicted by Figure 3. The case study design included characterization of research question, the units of analysis and the success criteria, as the main outcomes.

The PREDIQT-based analysis was performed by following the pre-defined process of the PREDIQT method. However, instead of performing predictions of effects of future changes during the last workshop (as specified by the PREDIQT process), we chose to demonstrate how prediction models can be applied by simulating the effects of reversal of a very large already implemented change. As such, the model application phase is not fully covered, but only demonstrated. The affected Design Model and DV elements were identified and their modified parameter values estimated by the domain experts. Thereafter, the simulation on the DVs was made by the analyst.

Additionally, in order to evaluate the predictions obtained, a thought experiment regarding the effect of the change on the root nodes of the respective DVs, was performed by the domain experts. Thus, this was a part of the method assessment. The overall assessment measures included: written feedback

(based on an evaluation template) from the analysis participants (affiliated with the customer organization) provided upon completion of the analysis and the above mentioned thought experiment based evaluation, verbal feedback during the analysis from the analysis participants (affiliated with the customer organization), and observations made by the analyst during the analysis. Based on the results of the PREDIQT-based analysis and the assessment, an evaluation with respect to the evaluation criteria was provided.

The research question of this case study is: *How does the PREDIQT method perform in a fully realistic setting and when applied on a system from a different domain than the previously evaluated one.*

PREDIQT phase-specific and the PREDIQT analysis general propositions are deduced. The propositions are then merged into a set of main success criteria and related to the objectives of the different stakeholders.

The units of analysis are identified as:

- the prediction models developed during the analysis
- the predictions obtained in terms of propagation paths and the QCF values
- the process of the PREDIQT-based analysis
- the participants of the analysis

The contents of the paper have been authored by the research group, which the analyst is a part of. In an attempt to avoid bias in the interpretation of the results, emphasis has been put on neutrally presenting the factual results, rather than interpreting and analyzing them in detail. The paper has been approved by the customer organization, with the aim of ensuring that agreement on the facts presented is achieved, as well as that no confidential information has been disclosed.

IV. Success criteria

Many concerns are relevant in evaluation of a method like PREDIQT. In order to efficiently cover most prevailing concerns, we start by identifying the stakeholder groups involved: the customers, the domain experts and the analyst. Success criteria (SC) are then deduced from the point of view of each stakeholder group. Note that the degree of relevance of each success criterion may vary between the stakeholder groups.

The *customers* are the ones needing, requesting and paying for the PREDIQT-based analysis. The customers are represented by decision makers, managers, system architects or personnel responsible for quality assurance. The customers are not necessarily involved in the the process of a PREDIQT analysis, but have interest in added value through enhanced decision making related to architectural design changes of the system and improved knowledge management in the organization. These two concerns should facilitate trustworthiness, reliability, usability and maintainability of the system. They should also decrease dependency of the organization on individuals with system or business critical knowledge.

For the customer of a PREDIQT-based analysis, the overall goal is to accomplish useful predictions. By useful predictions we mean predictions providing sufficient understanding of the impacts of the architectural design changes on system quality, so that informed decisions can be made. Hence,

SC1: *The PREDIQT-based analysis facilitates predictions providing sufficient understanding of the impacts of architectural design changes on system quality characteristics, so that informed decisions can be made.*

The customers' objective is also to be able to justify the cost of the analysis, compared to the benefit from it. Hence,

SC2: *The PREDIQT-based analysis is cost-effective.*

The *analyst* is the one conducting the PREDIQT analysis and documenting the results. This implies that the analyst has expertise on PREDIQT, leads the process, fully understands and in some cases participates in development of the prediction models, and documents the results. The analyst does however not necessarily have expertise on the target system under analysis, but should understand it sufficiently and be capable of collecting and processing the input needed in order to manage the development of the prediction models.

One objective for the analyst is to successfully conduct and document the analysis within the frame of the limited resources allocated. This implies that the PREDIQT-based analysis should be sufficiently simple to be feasible within the allocated resources, while still providing the requested added value for the customer. These goals are however already expressed through **SC1** and **SC2**. In addition, the analyst aims to capture through the prediction models the relevant knowledge, information and requirements on system architecture, usage profile, assumptions and constraints. This is crucial for ensuring quality of the model-based predictions in terms of both prediction certainty and the ability of the prediction models to handle the relevant architectural design changes. Hence,

SC3: *The prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality.*

The *domain experts* participate actively in all stages of the analysis. The domain experts may be represented by system architects, system users, developers, engineers, managers or experts in specific scientific areas that the system supports.

The PREDIQT method should help the domain experts communicating their knowledge in such a way that the analysis results in correct and sufficiently detailed prediction models which the participants agree upon and have a harmonized understanding of. The prediction models should therefore be comprehensible by the domain experts when properly guided by the analyst. Hence,

SC4: *The prediction models are sufficiently comprehensible to allow the domain experts to be actively involved in all phases of the PREDIQT process and achieve the goals of each phase with a common understanding of the results.*

Moreover, both the customer and the domain experts share the objective of improved knowledge management in the organization. For both, this is motivated by the concerns of efficient knowledge exchange, improved understanding of the system, reduced dependency on individuals, as well as increased maintainability and reliability of the system (as a

result of a model-based decision support). Hence,

SC5: *The PREDIQT-based analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality.*

V. Overview of the process undergone during the PREDIQT-based analysis

This section focuses on the process of the PREDIQT-based analysis (see Figure 3). We chronologically outline the relevant events and meetings in terms of their contents, participation, preparation and the time spent.

The analysis took place during the year 2010. Two preliminary meetings were held between the customer representatives and the analyst. The preliminary meetings were spent for motivating the analysis and identifying the challenges which the analysis should address. Thereafter, the analysis was organized in the form of five workshops and six working sessions in between some of the workshops. The workshops gathered both the domain experts and the customer (managers), and aimed to report on the current results and reach a milestone which the management should be involved in.The intermediate working sessions gathered the domain experts and the analyst to work tightly together on a particular task as a prerequisite for the forthcoming workshop. Table I outlines the process of the analysis. The first column specifies the type of the meeting (*PM*: preliminary meeting, *W*: workshop and *S*: working session) followed by the sequence number with respect to the kind of meeting. Column two specifies the date of the meeting. The third column lists the participants (note that all managers and domain experts are affiliated with the customer organization, while the analyst and the secretary belong to an external research group). The fourth column describes the contents and achievements of the meeting. The fifth column specifies the preparation activities for the meeting in question. The last column shows the approximate time spent (in terms of man-hours) during the meeting and in preparing for it. *T* denotes the total number of hours spent by all participants of the meeting (including the analyst), while *A* denotes the number of hours spent by the analyst only. The time spent on reporting and dissemination of the results after completion of meeting W5, is not included in the last column. At the end of each workshop (W), the approximate time until the next meeting was agreed upon, and the action points summarized. Minutes from each workshop were written and disseminated by the secretary. The sub-phases of the pre-defined process of the PREDIQT method (see Figure 1) were organized as follows:

- Target modeling
  - Characterize the target and the objectives: PM1, PM2, W1, S1
  - Create Quality Models: W2, S2, W3
  - Map Design Models: W2, S2, W3
  - Create DVs: S3, S4, S5
- Verification of the prediction models
  - Evaluation of prediction models: S5, W4, S6
  - Fitting of the prediction models: W4, S6
  - Approval of the final prediction models: S6
- Application of the prediction models
  - Specify the change: S6
  - Apply the change on the prediction models: W5
  - Quality prediction: W5

The case study was conducted in a realistic setting, with the objective of fully testing the feasibility of the method and providing added value for the customer. The target system analyzed serves within the customer organization as a semantic model and a repository for representation of the organization's core working processes, rules, and as a knowledge database. It is a business-critical expert system used for management and support of numerous working processes, involving hundreds of professional users every day. It is developed in-house, is rather complex and is used by numerous surrounding systems. The system represents an important asset for the customer organization. The changes on the system are implemented collectively approximately two times a year, while the individual changes are considered and designed frequently. The extent and number of changes are increasing. There is a requirement on the time to market of certain types of changes. The system and the associated semantic model are complex and it is therefore very hard to test all effects of changes (i.e. the cost of testing becomes an increasing problem). Alternative or complementing methods for testing are therefore desirable. For instance, prediction of change impacts can potentially be used to tune testing.

Different departments of the customer organization are responsible for the operation and the development of the system, respectively. The change planning and deployment (including adjustments and extensions) is, based on standardized procedures and tools, undertaken by domain experts of the department which is in charge of development of the system. The procedures for initiating, evaluating and carrying out the changes are well defined within the customer organization. The PREDIQT analysis is initiated by the organization's research department, on behalf of the overall stakeholders. Thus, the diversity of the participants and stakeholders (in terms of expertise, affiliation, interest, roles and background) is evident.

VI. Assessment

This section reports on the assessment part of the research method, depicted by Figure 3. Evaluation of the predictions based on a thought experiment is presented first. Secondly, the written feedback provided by the analysis participants from the customer organization upon completion of the above mentioned evaluation, is summarized. The written feedback is also referred to as a postmortem review. The third subsection reports on the verbal feedback provided, during the study, by the analysis participants from the customer organization. Lastly, the experiences and observations made by the analyst during the case study, are summarized.

TABLE I
OUTLINE OF THE PROCESS OF THE PREDIQT-BASED ANALYSIS

Meeting	Date	Participants	Contents	Preparation	Hours
PM1	March 25 2010	Two managers. The analyst.	Customer's presentation of the needs and challenges regarding quality, particularly security and interoperability of the systems. A brief presentation of the PREDIQT method and its possible application in the case study. Planning of the forthcoming meeting with the domain experts and the overall customer representatives.	Clarified formalities regarding communication channels and information exchange.	T:5 A:3
PM2	May 11 2010	Four managers. Three domain experts. The analyst.	Characterization (by the customer organization representatives) of the system architecture and main challenges that the case study may focus on. A presentation of the PREDIQT method and its possible application to the context.	The analyst received the input requested: system and enterprise architecture documentation, requirements specification, system design documentation, service level agreement and operational environment specification.	T:10 A:3
W1	June 15 2010	Three managers. Three domain experts. The analyst. The secretary.	The customer organization representatives characterized the target and the scope of the analysis: defined the target, defined the operational profile (current variability and expected changes in usage pattern, number of users, number of requests and amount of data), defined the expected lifetime of the system, specified type and extent of the expected changes, and characterized the main quality characteristics of the system.	The documentation studied by the analyst and clarifications or additional information needs communicated with the customer.	T:15 A:8
S1	June 17 2010	Two domain experts. The analyst.	Given to the analyst by the domain experts: a demo of the target system, a presentation of the functional properties of the system, specification of typical faults and failures due to changes of the system, and an overview of the testing procedures. Clarifications of the written input.	The analyst specified questions and additional information needs to the domain experts.	T:10 A:5
W2	Aug. 17 2010	Two domain experts. Three managers. The analyst. The secretary.	The analyst presented initial Quality Models (compliant with ISO 9126 [1]) and Design Models. Model revision in the group.	The analyst requested and received further documentation regarding system design. Development of system Quality Models and Design Models, by the analyst.	T:30 A:20
S2	Sept. 6 2010	Three domain experts. The analyst.	The analyst presented the updated Quality Models and Design Models. Selected use scenarios and change cases were undergone in the group, in order to check if the current models support their specification. Revision of all quality and Design Models in the group.	Updates (based on the discussion from W2 meeting) of system Quality Models and Design Models, by the analyst.	T:15 A:7
W3	Sept. 9 2010	Two domain experts. Three managers. The analyst. The secretary.	The analyst presented the current version of all prediction models. Revision of the Quality Models. Revision of the Design Models. Characterization of the types of potential architectural design changes. Preliminary approval of the available prediction models (Quality Models and Design Models).	Updates (based on the discussion from S2 meeting) of system Quality Models and Design Models, by the analyst.	T:20 A:10
S3	Sept. 28 2010	Four domain experts. The analyst.	The analyst presented the approach regarding the DV structure development (assumptions, rules, DV syntax and DV semantics) and an early draft of a DV, for the domain experts. Development of the DV structures in the group.	Development of an initial draft of a DV structure (by the analyst), for triggering the discussion and exemplification.	T:20 A:10
S4	Sept. 29 2010	Four domain experts. The analyst.	The analyst presented the approach regarding the (DV) parameter estimation (how to deduce the values, how to use the Quality Models, syntax and semantics of QCFs and EIs [16]), for the domain experts. Further development of the DV structures and DV parameter estimation in the group.	Documentation of the DV structure in the tool (MS Excel sheet customized for DVs in PREDIQT analysis). The analyst received documentation on typical system changes.	T:20 A:10
S5	Oct. 11 2010	Four domain experts. The analyst.	Further DV parameter estimation.	Documentation of the updated DVs in the tool.	T:15 A:5
W4	Oct. 20 2010	Three domain experts. One manager. The analyst. The secretary.	Validation of the DVs based on a thought experiment addressing randomly selected parts of the DVs. Model fitting of the DVs.	The analyst prepared a thought experiment setup based on the changes that the system has undergone.	T:20 A:8
S6	Oct. 22 2010	Two domain experts. The analyst. The secretary.	Continued validation of the DVs based on a thought experiment of addressing randomly selected parts of the DVs. Model fitting of the DVs. Final approval of the prediction models. Specification of changes which are to be simulated in the demo of meeting W5.		T:15 A:2
W5	Nov. 3 2010	Three domain experts. One manager The analyst. The secretary.	A summary of the results provided by the analyst: overview of the process undergone, and a presentation of the final prediction models. A demo of application of the prediction models: change specification, application of the change on the prediction models and quality prediction in terms of propagation paths and the modified QCF values.	The analyst prepared a simulation demo.	T:20 A:8

A. Evaluation of predictions

During the last part of the W5 meeting (that is, upon completion of the PREDIQT-based analysis), a thought experiment was performed by asking the domain experts to estimate the new root node QCF values on the respective DVs, due to a specified change (given the current and the new QCF values of the leaf nodes affected, as well as the current QCF value of the root node). The change specified was a major, already implemented architectural design change, which added a new functionality to the system. The evaluation (simulation and thought experiment) assumed reversal of the change. The change affected up to three leaf nodes on each DV. The purpose of the thought experiment was to test usefulness of the predictions obtained from the models. That is, we assume that the domain experts have thorough knowledge of the system, and that their root node estimates reflect the reality of how the quality characteristics are affected by the change. Then, the simulated root node value is compared to the thought experiment provided one. Since propagation during the simulation is subject to structure and parameter values of the prediction models, as well as the identified leaf nodes and their modified QCFs, all these aspects are incorporated into the evaluation when the simulated and the estimates (through the thought experiment) root node QCFs are compared.

The thought experiment showed the following relationship between the simulated root node QCF values and their corresponding estimates (provided by the domain experts), regarding the respective above presented simulations on:

- the first one of the two DVs dedicated to *Maintainability*: no deviation between estimated (by the domain experts) and simulated (by PREDIQT)
- the second one of the two DVs dedicated to *Maintainability*: estimated is 4.5% higher than simulated
- the first one of the two DVs dedicated to *Usability with respect to the contents*: estimated is 3% higher than simulated
- the second one of the two DVs dedicated to *Usability with respect to the contents*: estimated is 7.7% higher than simulated

B. Written feedback after the analysis

The summary provided here is based on contents analysis of the answers of five respondents. The answers have been provided on a pre-defined evaluation template (see Appendix 4). The answers have been abstracted and categorized in order to reduce the volume of raw text and reveal possible similarities and contrasts. More details from the written feedback are reported in Appendix 5.

All respondents have participated in the whole or parts of the analysis, and are affiliated with the customer organization. Table II summarizes the background of the respondents.

The main strengths pointed out are: "The PREDIQT method is useful and it suits well the problem addressed"(R2), "It was a way to in a systematic manner divide the problem in smaller parts, and then aggregate the quality level for the whole model"(R3), and "Modeling concept – propagation of assessments"(R4). A weakness repeatedly pointed out is the missing formal mapping of the parameter estimates to the model, i.e. the parameter estimates may be too sensitive to the context and the interpretation (R1, R3, R4, R5).

All five respondents agreed that the models facilitate communication, knowledge exchange and understanding of the target system, its architecture and its quality characteristics. R1 argues that "the workshops force people to communicate and harmonize into one model; the system is clarified and parts of the architecture are disclosed and discussed; the most important part is assigning estimates on quality characteristics, which forces people to make statements". R2 argues that "the method provides a good model of the system, which can be communicated around; when a multi-disciplinary group manages to make a model of a complex problem and communicate around it, you have achieved a good result; when you additionally can make predictions based on the model, the result is even better."

R1 points out that the effort needed for conducting the analysis is reasonable from a typical management consulting perspective, but in an engineering context, more effort should be directed towards specific parts.

Regarding the future use of the method, R1 expresses the intention to use the models developed in the future, for purpose of architecture development and dependability analysis. R2 and R3 express the wish to use the method in future projects, given that financing can be provided. R4 intends to use the prediction models if they can be tailored to specific use cases, while R5 writes: "I believe the model can be used to understand and predict the result/risk in different changes".

R1 expresses that the PREDIQT method "has already served the purpose in creating understanding and analysis. If incorporated with more tool support, I think it can be utilized in practice". R2 expresses that PREDIQT is very much better than no method, but it is unknown what it takes for it to be perfect. R3 and R4 express that the benefit from the method and quality of the predictions depend on the modeling skills and granularity of the models. R5 points out the challenge of interpreting the predictions due to the lack of documentation of the assumptions made during the parameter estimation.

Regarding challenges with usage of the method, R2 expresses two main issues: "access to competent resources to make the models and interpretation of the predictions and the corresponding uncertainty which requires competence". R3 points out three challenges: "be sure that you have modeled the most important aspects; models need to be verified; define the values in a consistent way". R4 sees the uncertainty challenge in the fact that the changes are marginal and therefore give small effects on the numbers, while R5 relates uncertainty to the insufficiently formal interpretation of the parameter values due to the assumptions made during their estimation.

Regarding the main benefit of the method, R2 expresses that PREDIQT "reduces uncertainty at change, but does not eliminate it; but it does systematize the uncertainty and reduce it sufficiently so that the method absolutely is valuable". R3 sees the discussion of the quality characteristics and agreement

TABLE II
BACKGROUND OF THE RESPONDENTS

	Respondent R1	Respondent R2	Respondent R3	Respondent R4	Respondent R5
Position	Senior Researcher	Chief Specialist	Software Architect	Senior Principal Engineer	Work Process Developer
Education (degree)	MSc	MSc	MSc equivalent	MSc	MSc
Years of professional experience	15	20	27	33	20
Role in the case study	Coordinator	Manager	Expert	Expert	Expert

upon the most important ones, as the main benefit.

The improvements suggested include simpler tool support, stricter workshops, increased traceability between the models, reuse of the Design Models based on other notations, and in-advance preparation of the experts.

C. Verbal feedback during the analysis

The verbal feedback includes the responses and comments from the analysis team, given during the different meetings – mainly by the end of the analysis. These include:

- The quality values (or their relative distance) should be mapped to monetary values or a similar measure of cost/gain in order to facilitate a cost-benefit analysis and ease interpretation of the DV parameters.
- The granularity of the changes is given by the granularity of the models. That is, minor changes may have very negligible impact on the models, unless the models are fine grained. A remedy is to deliberately increase the detail level of certain parts of the models. Still, although the parameters in such cases are almost unchanged, the prediction models help understand the propagation paths.
- The process of developing and verifying the models facilitates discussions, system understanding and knowledge exchange among the participants.
- The analyst should be aware of the possible bias or interests of the participants, particularly when the parameters are based on domain expert judgments.
- Certain parameters require a holistic approach (e.g. business perspective) or a special background (e.g. end-user). Some parameters may be uncertain due to lack of representation of such competence in the domain expert panel.
- Better documentation of the semantics and contextual information regarding the DV nodes, is needed. This would ease the use of DVs and particularly parameter estimation when some time has passed after the DV structure is developed.
- Active participation of the domain experts in the model development contributes not only to the model quality, but also to the experts' understanding of the models, and ability to use and maintain the models after the analysis.
- The time spent on development of the prediction models is much longer, than the time spent on the model verification. This has shown to be beneficiary, since model development was founded on numerous documents which the domain experts could interpret and relate to the quality notions. Doing this early in the process and consistently on all parts of the models while discussing the models in the group, is preferred to verifying certain parts of the models. Ideally, one should do both, but when the resources are limited, the choice we made was preferred (due to higher model quality early in the process, as well as more extensive brainstorming and discussions in the group) provided that the verification is satisfactory.
- The estimates are much more informative when considered and interpreted relative to each other, than individually. When one estimate is unambiguous in terms of the interpretation of the value and the assumptions made during its estimation, values of the others (on the same DV) may be compared to the well known one, in order to be interpreted.

D. Observations made during the analysis

Some of the main experiences and observations made by the analyst are presented in the sequel.

- One of the main challenges for the analyst during the development of the Design Models was acquiring an understanding of the expert terminology used in the system. The documentation received and the S1 meeting rectified this.
- Regardless of how well the analyst understands the target system and its quality characteristics, it is crucial that the analyst does not develop the prediction models alone. The model development and verification trigger many useful discussions among the domain experts, and help reveal inconsistencies and misunderstandings. In addition, the prediction models are intended to be used and maintained by the domain experts, who need to be able to relate to the models and the tools they are developed in. The optimal approach is that the analyst presents an initial version of the models, which are discussed, corrected and further developed in the group. Errors or missing parts in the initial models are often an advantage, as they trigger the discussions in the group.
- It is important to dedicate sufficient resources to characterization of the target, provision of the input and formation of a common understanding of the Quality Models. These are prerequisites for avoiding elementary discussions and ambiguities during the rest of the analysis.
- The analyst has to be aware of the inconsistencies of the terminology used in documents and the verbal communication among the domain experts, as well as between the overall stakeholders. Any such inconsistencies should

be clarified, preferably through the Quality Models or the Design Models.

- The PREDIQT method has to be sufficiently understood by all parties.
- It is important to use a notation for the prediction models, that all analysis participants can relate to.
- The time taken to estimate the parameters of the DVs is at least twice as long as the time needed to develop the structure of the DVs. It is necessary to explain that the DV structure is developed with respect to both Design Models and Quality Models, since dependencies are modeled with respect to the respective quality characteristic that the DV is dedicated to. Availability and common understanding of the Quality Models during parameter estimation is crucial.
- The structure of the DVs may need to be adjusted during the DV parameter estimation. For this, tool support more flexible than what our MS Excel sheets currently offer, is needed.
- When developing the DVs, certain assumptions and choices are made. Traces to the specific Design Model elements may exist, and only certain indicators from the Quality Models may be used in estimation. The current tool support is insufficient for efficiently documenting these aspects "on the run" during the meetings.
- Since a PREDIQT-based analysis requires considerable effort from the customer organization, it is essential to ensure commitment of the management and allocate the resources needed.
- It is important to make the right balance between the representativeness of the domain expert panel and the effectiveness of the analysis, when choosing the size of the analysis group. Although a larger group is likely to increase statistical significance of the data, time consumption on the discussions may rapidly grow with the number of the participants. Therefore, one should ensure that a fraction of the domain expert panel is present at all meetings and provides continuity, while some turnover of the overall participants depending on the goal of the meeting may be beneficiary. The turnover however necessitates updates of the participants on both the PREDIQT method and on the current status/results of the analysis. There is clearly a trade-off between the resource consumption and the model quality.
- The meetings should be as tightly scheduled as possible, provided that the necessary preparations are feasible. The rationale is to prevent the need to updates on recent results.
- Approximately half a year has been a reasonable time allocation for this case study. In a commercial analysis, a tighter course during a shorter period of time could be achieved, if the participants can prioritize the analysis even more among their overall tasks and if the tool support is improved.

VII. Evaluation with respect to the success criteria

In this section we evaluate the performance of the PREDIQT method in this case study, with respect to the success criteria presented in Section IV. Thus, this section addresses the last stage of the research method depicted by Figure 3.

SC1: *The PREDIQT-based analysis facilitates predictions providing sufficient understanding of the impacts of architectural design changes on system quality characteristics, so that informed decisions can be made.*

The ability of simulating a realistic change during meeting W5 and the assessment reported in Section VI, indicate that we have been able to develop an understandable and harmonized model of the system, communicate around the model, identify the dependencies and simulate the impacts of changes.

By performing thought experiments on the root node, the change propagation and its impact from the leaves throughout the different parts of the DVs, was evaluated. Whether the deviation reported is sufficiently small, is up to the customer to assess. The answers obtained in Section VI suggest that this is the case.

The thought experiment based evaluation of the predictions resulted in no deviation on the first DV, and some degree of overestimation during the thought experiments. This can be due to varying quality of the specific models or optimism of the domain experts. We observe however that the deviation between the simulated (based on the DV models) and the estimated (through the thought experiments) root node values during both model validation and the evaluation of the predictions, has no repeatable pattern but considerably high variance. Therefore, we do not have reason to assume bias in the relationship between the simulation and the thought experiments.

Many different parts of the DVs were affected during the evaluation, which ensured both variation and complexity in the change propagation – that is, coverage of the evaluation. Moreover, the number of parameters (QCFs and EIs) in each one of the four different DVs was around 60-70. Being able for a domain expert to remember the values and the structure of the four different DVs (which had been developed incrementally weeks before) should be improbable. Together with the above mentioned variance, this should exclude the possibility that the domain experts were able to quickly calculate propagation of the changes during the thought experiments. They were whatsoever asked by the analyst to purely use their system knowledge when performing the thought experiments.

The evaluation of the degree to which the simulated and the thought-experiment based estimates coincide would have been more reliable if uncertainty [19] had been expressed in the estimates. Then, one could have based the evaluation on whether the deviation is within the already present uncertainty of the estimates. Due to the limited time and the extent of the prediction models, we did not have the resources for also including the uncertainty handling in the analysis.

SC2: *The PREDIQT-based analysis is cost-effective.*

The analysis indicates that the PREDIQT method is feasible in a fully realistic setting and within the limited resources allocated. The process of the PREDIQT method was undergone, addressed the whole target of analysis and resulted in prediction models that, as the assessment indicates, provide the customer organization with useful basis for understanding the impacts of changes, capturing the propagation paths and obtaining the predictions. The domain experts have actively participated in development and revision of the Design Models and the Quality Models, and fully developed the DVs which cover the target of the analysis.

The feedback from R1 and R2 (customer management representatives) presented in Section VI, indicates cost-effectiveness of the analysis. The analysis has required approximately 215 (see Table I) man-hours (apart from the reporting), which is within the resources allocated. There are, however, some issues that must be taken into consideration when evaluating these numbers. Firstly, this was the second time the PREDIQT-based analysis was performed to a real industrial case. Hence, even though the analysis team included one of the inventors of the PREDIQT method, the process is not fully streamlined yet, due to limited empirical experience with PREDIQT. It can reasonably be assumed that the process will be more effective as the analysts gain experience with applying the PREDIQT method.

Furthermore, the process of the PREDIQT method assumes that the Design Models are in place prior to the analysis. Since this was not the case, considerable time had to be spent on modeling the system. Based on the experience gained and given that the Design Models are available as input to the analysis, we believe that it should be possible to carry out this kind of analysis within a time frame of approx. 60 man-hours spent by analyst (not including writing a final report) and ca. 50 man-hours spent by the overall participants. Hence, the success criterion appears to be fulfilled in this case. There is however still a need for a reference/baseline for comparing our results with the results from possible alternative methods. The future studies should address this, as well as cost-effectiveness per DV/quality characteristic/Design Model. Reusability of results (e.g. through experience factories) also contributes to the cost-effectiveness and should be examined in the future work.

SC3: *The prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality.*

The diversity of changes in the demo and the validation, the ability of simulating a realistic change during meeting W5 and the assessment, indicate that we have been able to develop a harmonized model of the system and use it for identifying the dependencies and simulating the impacts of all proposed changes. The participants provided a lot of information about the target during the analysis process. There were no instances where we were not able to capture the relevant information in the prediction models. Further application of the prediction models is however needed in order to evaluate their expressiveness and whether they can be maintained and used during the needed time period.

SC4: *The prediction models are sufficiently comprehensible to allow the domain experts to be actively involved in all phases of the PREDIQT process and achieve the goals of each phase with a common understanding of the results.*

The number of diagrams and parameter estimates was considerable. Still, the multi-disciplinary domain expert panel affiliated with several departments of the customer organization managed to discuss and agree upon the the different parts of the eventually harmonized and approved prediction models. The fact that the domain experts actively participated and continuously made progress according to the schedule of the analysis, managed to perform thought experiments and apply the models, indicates comprehensibility of the models. One of the most demanding parts of the analysis – development of the DVs, was entirely performed by the domain experts and only facilitated by the analyst.

The available prediction models were presented by the analyst during the meetings, in order to validate the correctness of the models or use them as basis for the forthcoming stages. There were many occasions where the participants suggested modifications, explained their rationale, or asked relevant questions about some detail in a model. This indicates that the models were in general comprehensible for the participants, and the postmortem review suggests that the models served well as an aid in establishing a common understanding of the target.

Still, comprehensibility of the models may vary among the participants and between the models depending on the knowledge of the system and the modeling notation. The fact that all the participants in this analysis had a strong technical background may have contributed to making the models easier for them to understand than would be the case for an even more diverse group. It is still necessary to have an analyst explain the method and the models, as well as facilitate and manage the process, since the current tool support is insufficient for ensuring a structured process and since an adequate PREDIQT manual currently does not exist. The analyst has played a rather active part during the analysis. A disadvantage is that the active role may have influenced the analysis. However, the involvement of the analyst is openly reported and reflected upon. It has also allowed better insight into the process and a more detailed evaluation of the results.

SC5: *The PREDIQT-based analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality.*

The answers reported in Section VI consistently suggest that the PREDIQT-based analysis facilitates knowledge management. The models have served as a means of documenting the system, triggering discussions and exchanging knowledge. The means of triggering the discussions and further increasing participation of the domain experts can still be developed as a part of the method. It is for example essential that the analyst does not too actively develop any models or uses the tools alone, which would make it more demanding for the domain experts to use and maintain the models.

More structured process, improved traceability between the models, documentation of assumptions and rationale, as well as improved tool support (in terms of flexibility of modifications, usability, process guidance, as well as documentation of traces, rationale and assumptions) would facilitate the knowledge exchange and certainty of the models.

VIII. Conclusions

The PREDIQT method makes use of models that capture the system design, the system quality notions and the interplay between system architecture and quality characteristics, respectively. The predictions result in propagation paths and the modified values of the parameters which express the quality characteristic fulfillment at the different abstraction levels. PREDIQT aims at establishing the right balance between the practical usability of the models, and the usefulness of the predictions. We are not aware of other approaches that combine notions of architectural design and quality in this way. However, the issues of metrics estimation, system quality and the various notations for modeling system architecture, have received much attention in the literature [8, 15, 11, 1, 3, 2, 9, 7, 13, 12, 4, 6, 14].

The paper has presented experiences from using the PREDIQT method in an industrial case study. The contributions of the paper include:

1) a detailed account of how the PREDIQT method [16] scales in an industrial context
2) an evaluation of the performance of the method in an industrial context.

The experiences and results obtained indicate that the PREDIQT method can be carried out with limited resources (five workshops and 215 man-hours), on a real-life system and result in useful prediction models. Furthermore, the observations indicate that the method, particularly its process, facilitates understanding of the system architecture and its quality characteristics, and contributes to structured knowledge management through system modeling. All stakeholders, including the customer, the domain experts and the analyst gained a better and a more harmonized understanding of the target system and its quality characteristics, during the process. The knowledge management in the context of this case study has concerned acquisition, exchange and documentation of the knowledge available (in forms such as domain expert knowledge, documentation or logs), regarding the architectural design of the system, non-functional (quality) characteristics of the system and the interplay between the architectural design and the system quality.

Four evaluation methods have been used: a thought experiment in order to evaluate the predictions obtained; written feedback from the analysis participants; verbal feedback during the analysis from the analysis participants; and observations made by the analyst during the case study. The evaluation methods complement each other and are to a varying degree used during the discussion of the success criteria. For example, when discussing success criterion 1 (usefulness of the predictions for making informed decisions), the thought experiment is mainly referred to.

The issue of method scalability concerns two aspects which our results indicate have been achieved and balanced: resources required to perform the analysis and the usefulness of the prediction models. In particular, the evaluation argues that:

- the PREDIQT-based analysis facilitates predictions providing sufficient understanding of the impacts of architectural design changes on system quality characteristics, so that informed decisions can be made,
- the PREDIQT-based analysis is cost-effective,
- the prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality,
- the prediction models are sufficiently comprehensible to allow the domain experts to be actively involved in all phases of the PREDIQT process and achieve the goals of each phase with a common understanding of the results, and
- the PREDIQT-based analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality

within the scope of the characterized target system and objectives.

Full documentation of the case study exists, but its availability is restricted due to confidentiality required by the customer. Hard evidence in the form of measurements to validate the correctness of the predictions would have been desirable, but this was unfortunately impossible within the frame of this case study. Instead, we have relied on extensive documentation and the domain expert group with solid background and diversity. Still, thought experiment based validation of models and evaluation of the predictions have weaknesses compared to the measurement based ones. Particularly, we can not exclude that possible undocumented or inconsistent assumptions have been made in model development, although the Quality Models and the active participation of the domain experts in all model development should prevent this. Statistical power was limited, due to low number of participants. The careful selection of experienced participants and the variety of the changes specified during model validation, compensated for some of this. Another weakness is that the same domain expert group has developed and validated the prediction models. However, given the complexity of the prediction models (which are very unlikely to be remembered), the variation of the changes applied and variance of the deviation pattern obtained (between the simulations and the thought experiment based estimates), we can not see any indication of bias due to the same expert group.

Although the above mentioned threats to validity and reliability are present in such a study, we argue that the results indicate the feasibility and usefulness of the method in a real-life setting. The study has also provided useful insight into the current strengths and weaknesses of the method, as well as suggested directions for future research and improvements.

Particularly, the needs for improved traceability, even more structured process guidelines and better tool support have been highlighted.

Note that PREDIQT has only architectural design as the independent variable – the Quality Model itself is, once developed, assumed to remain unchanged. This is of course a simplification, since system quality prediction is subject to more factors than architectural design. Usage profile, quality definitions and process are examples of the factors whose variation PREDIQT does not address. Although this case study has evaluated PREDIQT in a different domain compared to the one reported in [16], many more evaluations are needed for evaluating the external validity of the method.

The target system is representative for the systems intended to be within the scope of the PREDIQT method. This is the second trial of PREDIQT in a real-life setting and both trials have given strong indications of feasibility of the method, reported similar benefits (understanding of system architecture and its quality, usefulness of estimates particularly when interpreted relative to each other, and usefulness of the process) and undergone the same stages of the PREDIQT process. There is no significant difference in the size or complexity of the prediction models between the two case studies. No particular customizations of the method were needed for this trial. Thus, we have reason to believe that it should be possible to reapply PREDIQT in another context.

Acknowledgments: This work has been conducted as a part of the DIGIT (180052/S10) project funded by the Research Council of Norway, as well as a part of the SecureChange project and the NESSoS network of excellence both funded by the European Commission within the 7th Framework Programme.

References

[1] International Organisation for Standardisation: ISO/IEC 9126 - Software engineering – Product quality. 2004.

[2] V. Basili, G. Caldiera, and H. Rombach. The Goal Question Metric Approach. *Encyclopedia of Software Engineering*, 1994.

[3] V. R. Basili. Software Modeling and Measurement: the Goal/Question/Metric Paradigm. Technical Report TR-92-96, University of Maryland, 1992.

[4] C. Byrnes and I. Kyratzoglou. Applying Architecture Tradeoff Assessment Method (ATAM) as Part of Formal Software Architecture Review. Technical Report 07-0094, The MITRE Corporation, 2007.

[5] T. D. Cook and D. T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Houghton Mifflin Company, 1979.

[6] L. Dobrica and E. Niemel. A Survey on Software Architecture Analysis Methods. *IEEE Transactions on Software Engineering*, 28(7):638–653, 2002.

[7] C. Ebert, R. Dumke, M. Bundschuh, A. Schmietendorf, and R. Dumke. *Best Practices in Software Measurement.* Springer Verlag, 2004.

[8] N. Fenton and M. Neil. A Critique of Software Defect Prediction Models. *IEEE Transactions on Software Engineering*, 25:675–689, 1999.

[9] N. E. Fenton and S. L. Pfleeger. *Software Metrics: A Rigorous and Practical Approach.* PWS Publishing Co., 1998.

[10] B. Flyvbjerg. Five misunderstandings about case-study research. *Qualitative Inquiry*, 12(2):219–245, 2006.

[11] D. Heckerman, A. Mamdani, and W. M. P. Real-World Applications of Bayesian Networks. *ACM Communications*, 38(3):24–26, 1995.

[12] R. Kazman, M. Barbacci, M. Klein, S. J. Carriere, and S. G. Woods. Experience with Performing Architecture Tradeoff Analysis. *International Conference on Software Engineering*, 0:54, 1999.

[13] R. Kazman, M. Klein, M. Barbacci, T. Longstaff, H. Lipson, and J. Carriere. The Architecture Tradeoff Analysis Method. In *Fourth IEEE International Conference on Engineering of Complex Computer Systems*, pages 68–78, Aug 1998.

[14] M. Mattsson, H. Grahn, and F. Mårtensson. Software Architecture Evaluation Methods for Performance, Maintainability, Testability and Portability. In *Second International Conference on the Quality of Software Architectures*, June 2006.

[15] M. Neil, N. Fenton, and L. Nielsen. Building Large-Scale Bayesian Networks. *Knowledge Engineering Rev.*, 15(3):257–284, 2000.

[16] A. Omerovic, A. Andresen, H. Grindheim, P. Myrseth, A. Refsdal, K. Stølen, and J. Ølnes. A Feasibility Study in Model Based Prediction of Impact of Changes on System Quality. Technical Report A13339, SINTEF, 2010.

[17] A. Omerovic, A. Andresen, H. Grindheim, P. Myrseth, A. Refsdal, K. Stølen, and J. Ølnes. A Feasibility Study in Model Based Prediction of Impact of Changes on System Quality. In *International Symposium on Engineering Secure Software and Systems*, volume LNCS 5965, pages 231–240. Springer, 2010.

[18] A. Omerovic and K. Stølen. Simplifying Parametrization of Bayesian Networks in Prediction of System Quality. In *Third IEEE International Conference on Secure Software Integration and Reliability Improvement*, pages 447–448. IEEE, 2009.

[19] A. Omerovic and K. Stølen. Interval-Based Uncertainty Handling in Model-Based Prediction of System Quality. volume 0, pages 99–108. IEEE Computer Society, 2010.

[20] W. R. Shadish, T. D. Cook, and D. T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Houghton Mifflin Company, 2001.

[21] R. K. Yin. *Case Study Research: Design and Methods, Third Edition, Applied Social Research Methods Series, Vol 5.* Sage Publications, Inc, 3 edition, 2002.

APPENDIX 1: RESEARCH METHOD

The research method is motivated by the guidelines for case study research provided by Yin [21]. This section reports on the rationale for main decisions regarding case study design, data acquisition and the analysis. A deductive approach is undertaken, where the already defined PREDIQT method is exposed to an empirical trial in the form of a case study. The structure of this section follows the structure from [21].

A case study

The technical definition of a case study is as follows [21].

1) *A case study is en empirical inquiry that:*
   - *investigates a contemporary phenomenon within its real-life context, especially when*
   - *the boundaries between phenomenon and context are not clearly evident.*
2) *The case study inquiry:*
   - *copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result*
   - *relies on multiple sources of evidence, with data needing to converge in a triangulating fashion, and as another result*
   - *benefits from the prior development of theoretical propositions to guide data collections and analysis.*

A case study method is, according to [21], used when the researcher deliberately wants to uncover contextual conditions – believing that they might be highly pertinent to the phenomenon of study. A case study comprises an all-encompassing method – covering the logic of design, data collection techniques, and specific approaches to data analysis.

The main stages of the research method performed in this case study are depicted by Figure 3. The case study design included characterization of the research question, the units of analysis and the success criteria, as the main outcomes.

The PREDIQT-based analysis was performed by following the pre-defined process of the PREDIQT method. However, instead of performing predictions of effects of future changes during the last workshop (as specified by the PREDIQT process), we chose to demonstrate how prediction models can be applied by simulating the effects of reversal of a very large already implemented change. As such, the model application phase is not fully covered, but only demonstrated. The affected Design Model and DV elements were identified and their modified parameter values estimated by the domain experts. Thereafter, the simulation on the DVs was made by the analyst.

Additionally, in order to evaluate the predictions obtained, a thought experiment regarding the effect of the change on the root nodes of the respective DVs, was performed by the domain experts. Thus, this was a part of the method assessment. Besides the thought experiment, the assessment measures included: written feedback (based on an evaluation template) from the analysis participants (affiliated with the customer organization) provided upon completion of the analysis and the above mentioned thought experiment based evaluation; verbal feedback during the analysis from the analysis participants (affiliated with the customer organization); and observations made by the analyst during the analysis. Based on the results of the PREDIQT-based analysis and the assessment, we provide an evaluation with respect to the evaluation criteria.

Design of the case study

The research question of this case study is "How does the PREDIQT method perform in a fully realistic setting and when applied on a system from a different domain than the previously evaluated one". The propositions are deduced based on the three phases of the process of the PREDIQT method. Since the earlier performed case study [16] indicated feasibility of the PREDIQT method, the objective of this one is to investigate the performance of the method in a more structured manner. Furthermore, as proposed by [21], we included the following theory in the case study design:

The case study will show the weaknesses, the strengths and the cost-effectiveness of the PREDIQT method when exposed in a new domain, as well as within a fully realistic and organizationally complex setting. This will provide the insight into feasibility and performance of the method. Particularly we will uncover if the method still scales in terms of resource consumption and the size of the prediction models, when performed on a different domain and in a more complex setting compared to the earlier case study. The case study will also identify the needs for further research.

Phase 3: "Application of prediction models" is successful if:

- the prediction models are sufficiently expressive to adopt the relevant architectural design changes and analyze their effects on quality
- the prediction models can indicate the change propagation paths and provide the resulting quantitative values (of the quality characteristics) with sufficient certainty
- the prediction models are sufficiently comprehensible to the domain experts so that the specified changes can be applied and the predictions (propagation paths and the modified quality characteristic values) can be interpreted.

Phase 2: "Verification of prediction models" is successful if:

- the prediction models are sufficiently comprehensible to the domain experts so that they can be evaluated, fitted and approved with a common understanding
- the prediction models are approved by the domain expert panel with regard to their expressiveness, granularity, correctness and completeness in representing the target system within the objectives of the analysis. That is, the prediction models include the necessary representation of the target system and can adopt the kinds of changes characterized in Sub-Phase 1 of Phase 1.
- the prediction models are approved by the domain expert panel with regard to certainty of the estimates.

Phase 1: "Target modeling" is successful if:

- the process and the objectives of the PREDIQT analysis are comprehensible to the domain experts so that a

correct characterization of the objectives and input can be provided
- the prediction models represent the target system within the objectives of the analysis
- the prediction models are sufficiently comprehensible to the domain experts so that they can be actively involved in their development and achieve a common understanding and agreement with respect to the models.

Additionally, a PREDIQT-based analysis is successful if:

- the analysis is cost-effective
- the analysis facilitates knowledge management and contributes to a common understanding of the target system and its quality.

The phase-specific propositions and the entire analysis related ones are related to the objectives of the different stakeholders and merged into a set of main success criteria, as presented in Section IV.

Based on the propositions, the units of analysis are identified as:

- the prediction models developed during the analysis
- the predictions obtained in terms of propagation paths and the QCF values
- the process of the PREDIQT-based analysis
- the participants of the analysis

The observations, the data collection and the postmortem review are therefore focused on the four specified units of analysis. Since reporting of all experiences from such a broad case study is rather inexpedient, this paper reports on the aspects relevant for the criteria and units of analysis specified above.

The evaluation of the research suggested by [21] is based on four tests: construct validity, internal validity, external validity and reliability.

The three tactics available by [21] for increasing construct validity applied in this case study are:

- use of multiple sources of evidence; in our context: extensive documentation, analyst's observations, multiple and diverse change specifications, expert judgments and written evaluation from a diverse group of domain experts
- establish a chain of evidence; in our context: the analyst has documented and archived all models, meeting notes and presentation slides, and the secretary has documented and archived minutes from each workshop and agenda of each workshop. All documentation provided, evaluation forms filled out and emails exchanged have also been archived. All reported events, procedures, artifacts and participants and parties involved are traceable by the authors. The confidentiality demanded by the authors prevents however full traceability by the external parties.
- have a draft case study report reviewed by key informants; this is fulfilled through a quality assurance of this report by the case study coordinator (a management representative from the customer organization). The case study coordinator participated in all preliminary meetings and workshops, followed up the progress and served as the communication point at customer side.

The four analytic tactics available by [21] for increasing internal validity applied in this case study are:

- do pattern matching; this is done through comparing the simulated to the thought experiment based effects of the changes applied on one or more leaf nodes (independent variables), on the root node (dependent variable) of the DV. The procedure is applied on different leaf nodes of each DV, and based on multiple independent changes. Moreover, the changes have previously been implemented and their effects are known (they have therefore been reversed in the prediction models). Additionally, a diverse group of the domain experts has been involved in performing the thought experiments. A weakness is however that the same domain expert group has developed the prediction models. However, given the complexity of the prediction models which is unlikely to be remembered, the variation of the changes applied and variance of the pattern obtained, we can not see any indication of bias due to the same expert group.
- do explanation building; the deviations obtained from comparing the simulation results to the thought experiment based estimates are explained through model quality and expert optimism. The data collected through the evaluation is however not representative for substantiating the explanation.
- address rival explanations; the rival explanation suggesting a bias due to memorization of the models by the experts, is rejected due to complexity and inaccessibility of the DVs.
- use logic models; this related the quality characteristic decompositions to the estimates of the leaf nodes on the respective DVs, and their further propagation to the root nodes. The above mentioned measures from the pattern matching apply.

One of the two tactics available for increasing external validity, is applied in this case study:

- use theory in single-case studies; the theory is specified above.
- use replication logic in multiple-case studies; only partially applied, this being the second trial of PREDIQT in an industrial setting.

Two tactics are available for increasing reliability of a case study:

- use case study protocol
- develop case study database

As mentioned above, full documentation of the case study exists, but its availability is restricted due to confidentiality required by the customer.

Due to several units of analysis and fulfillment of one type of rationale for single-case study design, our case study is classified as embedded single-case study. The types of rationale for a single-case study are:

- critical case; we are testing an earlier prescribed method with the aim to confirm, challenge or extend it. However, since a critical case should be designed so that it can be used to generalize or falsify a theory [10], we can not claim that our case is critical.
- testing a unique or an extreme case; neither unit of analysis is considerer to be extreme or unique, given the frames of a typical analysis.
- representative or typical case; we believe to have captured circumstances and conditions of a realistic case study.
- revelatory case; we do not consider the units of analysis to previously having been inaccessible for scientific investigation.
- longitudinal case; this case has not been studied at several points of time.

Presence of one rationale is, according to [21], sufficient for choosing a single-case study. The embedded single-case study allows focusing on specific units of measure, while enabling study of the rest of the context. Thus, the evaluation is targeted, while the larger units are included as well.

The sample size and the data quality have been given by the resources available within the study. The extent of the documentation, number of participants, qualifications of the participants and the resampling effort have defined this. We have fully utilized all resources available within the frames of the analysis.

Preparing data collection

Yin [21] emphasizes the skills of the analyst an an important prerequisite. Given the analyst's professional background and the role in development and earlier trial of the PREDIQT method, we consider this condition to be fulfilled.

Another prerequisite is the training and preparation for the case study. Given the preliminary meetings when the method was presented, as well as systematic guidance of the analysis participants provided throughout the case study, we consider this condition to be fulfilled. The goal has been to have all participants agree upon the objectives, understand the basic concepts, terminology, the process, the rationale and the issues relevant to the study. Discussions rather than presentations have been the key approach, in order to ensure that the desired level of understanding has been achieved.

The third prerequisite is the protocol development. The meeting notes, minutes from the workshops, meeting presentation slides and all models have been either produced in groups, or reviewed once they are produced.

The fourth prerequisite – screening of the case study nominations involved nomination of the participants of the analysis and characterization the the target system. The respective nominations were done by the customer organization and the established analysis team, respectively.

Collecting the evidence

Yin [21] discusses six sources of evidence: documentation, archival records, interviews, direct observation, participant-observation, and physical artifacts which are highly complementary to each other. Additionally, [21] presents three essential data collection principles:

- use multiple sources of evidence
- create a case study database
- maintain a chain of evidence

The documentation has included: administrative documents, minutes from meetings, presentation slides, meeting notes, filled evaluation forms, system and enterprise architecture documentation, requirements specification, system design documentation, service level agreement, operational environment specification, procedural descriptions for change request in the organization, information model of the system, and the prediction models developed.

The archival records are participant contact details, the disclosure agreements, e-mail correspondence, and a listing of challenges from the preliminary meetings.

The interview form applied was structured conversation with the analysis participants. The statements have been documented in form of models and meeting notes.

The analyst has reported on the main direct observations and participant observations from the case study. The analyst has played a rather active part during the analysis. A disadvantage is that the active role may have influenced the analysis. However, the involvement of the analyst is openly reported and reflected upon. It has also allowed better insight into the process and a more detailed evaluation of the results. Still, the target characterization, the model revisions and approvals and the evaluations have purely been performed by the domain experts who were only guided by the analyst. Additionally, the participant observations from the overall participants have been collected in written and verbal forms, and reported as a part of the postmortem review.

The multiple sources of evidence used in the case study have developed *converging lines of inquiry*, a process of triangulation which makes a conclusion more valid when based on several sources of correlating evidence. Documentation, informers, data, responders and observers have been triangulated. The triangulation is a means of increasing the construct validity.

As mentioned above, all documentation is stored and traceable, but its availability is restricted due to the confidentiality. The models have been stored in their incremental versions. The documentation as such provides a chain of evidence which allows tracing the origins of the main results.

Analyzing case study evidence

The case study has relied on the research question stated above, and the evaluation has been driven by the success criteria specified in Section IV. The success criteria serve as propositions. The evaluation template for postmortem also addresses the main success criteria. The rival explanations have been considered as a part of the section which discusses threats to validity and reliability, but due to inability of determining statistical significance of the input, no null hypothesis has been formulated. The prediction models have been analyzed in terms of their size, complexity, comprehensibility and the

deviation between the simulated and the thought experiment based estimates. Furthermore, the results of the postmortem review have been summarized based on contents analysis of the answers which are abstracted and categorized in order to reduce the volume of raw text and reveal possible similarities and contrasts.

As mentioned above, pattern matching has been performed during both model validation and model application, by comparing the simulated estimates with the ones obtained through thought experiments. The patterns are related to the dependent variables (root nodes) but they also validate the affected parts of the DV which are involved in the propagation from the modified leaf nodes. If the patterns coincide sufficiently, it helps strengthen the internal validity.

The validation is based on multiple independent changes which address different parts of each DV. Moreover, a multi-disciplinary expert panel was involved in the thought-experiment based evaluation. The change which the final evaluation was based on was realistic, extensive, known and affected several parts of each DV. Both propagation paths and the values obtained from the simulation, were evaluated.

Uncertainty handling was not included due to the limited resources. The priority was rather to develop the prediction models which cover the target of the analysis. However, extensive use of the Design Models and other documentation, as well as discussions when developing the DVs did aim at increasing the precision.

Comparison of two case studies and cross-case synthesis is another means of explanation building. This case will be briefly compared to the previous one, in relation to the discussion of the threats to external validity.

Logic models are presented in [21] as a technique for stipulating complex chain of events over time. In a logic model, the dependencies are modeled and the data acquisition is planned for testing the effects of changes of the modified parts on the related ones. In our case, the DVs and the evaluation addressing the root nodes. The structure of the Quality Models is also indirectly tested, as the estimates are based on it. The DV structure is also based on the quality characteristic definition, as dependencies are also expressed with respect to the quality characteristic. The DVs are also partially traceable elements of the Design Models. Thus, the relationships between the prediction models, as well as relationships between the nodes of a DV may be considered as a logic model.

Reporting the case study

The contents of this report have been driven by the success criteria and the related reporting needs. The audience is the research community and the practitioners interested in future trials of the PREDIQT method. [21] provides guidelines on the reporting of a case study. One of the issues addressed is the anonymity, which is accepted when absolutely necessary. Due to the confidentiality requested by the customer, the concrete context and the results have only been reported to the degree approved. Still, we believe that the paper provides useful insight into the experiences from the trial, and fulfills the objective regarding evaluation of the method, as well as suggestion of the future work.

The contents of the paper have been authored by the research group, which the analyst is a part of. In an attempt to avoid bias in the interpretation of the results, emphasis has been put on neutrally presenting the factual results, rather than interpreting and analyzing them in detail. The relevant results have been presented, including both supporting and challenging data. The selectiveness is, as argued by [21], relevant in limiting the paper to the most critical evidence, instead of cluttering the presentation with supportive but secondary evidence (which may sway or bore the reader). The paper has been approved by the customer organization, with the aim of ensuring that agreement on the facts presented is achieved, as well as that no confidential information has been disclosed. [21] argues that such an approval increases the construct validity of a study.

Appendix 2: Setup and data collection during the PREDIQT-based analysis

The analyst had more than nine years of relevant professional experience in software engineering. The customer management representatives and the domain experts had between 15 and 33 years of relevant professional experience each.

The system documentation received by the analyst from the customer organization contained mostly descriptions and specifications in the form of verbal input, presentation slides, textual documents, sketch-like models integrated in various documents, samples of change request forms, MS Excel documents in which the system structure is described, and webpages. Thus, the Design Models of the target system had to be developed as a part of the analysis. All Design Models and Quality Models were developed in UML, using the MS Visio tool. The DV structure was developed in MS Visio, and then transferred to an already developed MS Excel based tool [16]. The latter tool displays DVs and supports automatized parameter propagation and sensitivity analysis.

The original input and presentations provided in relation to the meetings PM1 and PM2 were, in terms of scope of the systems presented and abstraction level, considerably broader than the target of the analysis defined at the meeting W1. During the W1 meeting, the group was guided by the analyst (using precise questions and examples) to characterize the target in terms of scope and the quality characteristics. This input, in addition to the written documentation already received, was sufficient for the analyst's development of the initial Design Models and Quality Models of the system. The quality characteristics identified during W1 were compiled to the relevant parts of the ISO 9126 [1] standard, where the interpretations and formal definitions of all elements are provided. All quality characteristics in the compiled model were decomposed into sub characteristics and indicators, in accordance with the relevant parts of the ISO 9126 standard.

Because verification only addresses certain parts of the models, it was deliberately chosen to spend more resources

on the model development, particularly estimation of the DV parameters, than on their verification. Much of the existing documentation was used to deduce the values, rather than to verify them afterwards.

All DVs were (in terms of structure and parameter estimations) entirely developed and revised by the domain expert panel. Support from the analyst included guidance on the procedures, rules, model syntax and documentation.

At meeting W4 the focus was on validation. The objective was to check whether the models can predict within an acceptable threshold, in which case they are approved. After a walkthrough of the four DVs, the following procedure was followed for 9 independent change simulations:

1) the analyst presented the leaf node in question, its parent node, the DV it is a part of and the current QCF of the leaf node in question, without showing the DV
2) the domain experts specified a known change which has already been implemented on the system and which influenced the selected leaf node
3) the analyst presented the current QCF of the root node
4) the analyst asked the domain expert panel for an estimate of the new QCF of the node in question, given that the change is deployed on the current system
5) the analyst asked the domain expert panel for an estimate of the new QCF of the root node
6) the analyst used the DV to predict the new root node value after the change.

On the two *maintainability* DVs, 2 and 3 leaf nodes were modified, respectively. On the two *usability with respect to contents* DVs, 2 different leaf nodes on each DV were modified. Each node was modified independently according to the procedure above. In case the change specified already has been applied on the system, its reversal is estimated and simulated during the validation. A table with the parameters from the procedure was dedicated to each change and displayed on the presentation slides.

At meeting W5 the focus was on demonstrating application of the prediction models. The demo included:

1) Specification of a change which has already taken place
2) Identification of the Design Model elements affected (By domain experts))
   - Specification and substantiating of the Design Model changes
3) Identification of the related parts of the DVs (By domain experts))
   - Modification of the DV parameter values
4) Simulation of the change propagation on the DVs (By analyst)
5) Documentation of the DV nodes affected and the modified QCF values

The demo of model application assumed reversing a major change which has already been deployed on the system. Hence, the current prediction models incorporated the state after the change whose reversal was to be demonstrated.

At each meeting, handouts of the current prediction models and presentation slides were provided to all participants.

APPENDIX 3: OUTCOMES OF THE PREDIQT-BASED ANALYSIS

This section reports on the main outcomes of the process presented in the previous section. We focus particularly on the final prediction models, and the result of their validation and application carried out during the respective stages of the process.

Characteristics of the prediction models

Both Quality Models and Design Models were developed using UML. The Quality Models decomposed the total quality of the system into two main quality characteristics: *maintainability* and *usability with respect to contents*. *Maintainability* was decomposed into three sub-characteristics (changeability, testability and stability), which again were decomposed into three, three and two indicators, respectively. *Usability with respect to contents* was decomposed into three sub-characteristics (information correctness, information availability and operability), which again were decomposed into four, four and three indicators, respectively. In addition, each subtree was supplemented by a node called "Other", for model completeness purpose. All nodes of the Quality Model (except the ones called "Other") were defined qualitatively and formally. Most of the definitions were retrieved from the ISO 9126 standard, the remaining ones were customized with respect to the target system. All formal definitions ensured normalized values between zero and one, where 0 denotes no fulfillment, and 1 maximum fulfillment.

The Design Models specified concepts, system structure and workflow. The Design Models consisted of 10 UML diagrams. Mostly, class diagrams were used and their size ranged over 10-20 classes. One activity diagram specified the workflow, and contained 10 (activity) elements and one decision point.

Two DVs were developed for each quality characteristic, that is, four DVs in total. The two DVs dedicated to a quality characteristic covered the two main perspectives of the system's purpose. The same two perspectives were covered by the remaining two DVs for the other quality characteristic, but (in terms of structure) the DVs of the second quality characteristic were to a limited degree different from the respective corresponding DVs dedicated to the first quality characteristic. The main difference was in names and semantics of certain nodes. The parameter values are not comparable between the DVs of two quality characteristics.

For the quality characteristic *Maintainability*, the first DV had 31 nodes, of which 26 nodes were leaves. The second DV had 35 nodes, of which 28 nodes were leaves. For the quality characteristic *Usability with respect to contents*, the first DV had 31 nodes, of which 26 nodes were leaves. The second DV had 34 nodes, of which 27 nodes were leaves

Results of the model validation

Table III summarizes the results of the above presented validation. The first column specifies the DV on which the

change is introduced. The second column shows the difference between the estimated (i.e., modified due to the change by the domain expert panel) and the old (prior to the change) value of QCF on the leaf node addressed. The third column shows the difference between the simulated (by the DV) and the old (prior to the change) value of QCF on the root node of the DV. The last column shows the difference between the simulated (by the DV) and the estimated (by the domain experts) value of QCF on the root node of the DV.

Results of the demonstrated application of the prediction models

Of the 10 existing diagrams of the Design Models, 7 diagrams were affected by the change specified. Within these 7 diagrams, the number of elements affected by the change was 2 out of 4, 1 out of 20, 5 out of 8, 2 out of 9, 1 out of 6, 2 out of 12 and 2 out of 5, respectively.

On the first one of the two DVs dedicated to *Maintainability*, three leaf nodes were affected as follows: increase by 1%, increase by 1% and unchanged QCF, respectively. The simulation showed that one internal node was affected and no significant effect on the root node was indicated by the simulation.

On the second one of the two DVs dedicated to *Maintainability*, one leaf node was affected by a decrease of 30%. The simulation showed that one internal node was affected and the simulated effect on the root node was a decrease of QCF by 3%.

On the first one of the two DVs dedicated to *Usability with respect to the contents*, QCFs of three leaf nodes were affected as follows: decrease by 3%, decrease by 2.5% and decrease by 8%, respectively. The simulation showed that one internal node was affected and the simulated effect on the root node was a decrease of QCF by 3%.

On the second one of the two DVs dedicated to *Usability with respect to the contents*, one leaf node was affected by a decrease of 30%. The simulation showed that one internal node was affected and the simulated effect on the root node was a decrease of QCF by 3%.

APPENDIX 4: DESIGN OF THE EVALUATION TEMPLATE

The evaluation template, used in relation to the postmortem review, was designed in the form of a questionnaire in MS Word form, as follows:

Title: *Evaluation of the PREDIQT method in the XXX case study*

Introduction *We need your feedback in order to further improve the PREDIQT method. Can you please provide your answers and comments to the following questions? All questions are regarding the case study you have been involved in during the second half of the year 2010, that is: "Analysis of XXX system, based on the PREDIQT Method".*

DATE:

1. Please specify your background and role in the case study:
   - Work place:
   - Position:
   - Education (degree):
   - Years of professional experience:
   - Role in the case study:

2. What is your general impression of the PREDIQT method? Please describe the experience from the case study in your own words. What do you think are strengths and weaknesses of the PREDIQT method? You may comment on both the process undergone and the final prediction models.

3. To what degree do you think the method (including the process and the final prediction models) facilitates communication, knowledge exchange and understanding with respect to:
   - the XXX system in general,
   - its architecture, and
   - its quality characteristics?

4. What is your experience from the process undergone? We are particularly interested in your opinion regarding the effort needed to develop the final prediction models, your understanding of the process and your opinion on the involvement or interest of the different participants, during the case study.

5. Please comment on your opinion regarding certainty, understandability, completeness and usability of the prediction models:
   - Design Models
   - Quality Models
   - Dependency Views

6. Do you intend to make use of any prediction models (Design Models/Quality Models/DVs), in the future? If so, which models do you intend to use further and in which context? If not, why not?

7. To what degree do you think the PREDIQT method (the process undergone and the resulting models) can aid understanding, analyzing and predicting the impacts of changes of XXX on its quality?

8. What do you see as the main challenges or problems with usage of the method and the final prediction models?

9. What do you see as the main benefits of the method (process and the prediction models)?

10. Which prediction models (Design Models/Quality Models/DVs) or properties of the models do you find most useful and why?

11. What kinds of improvements of the PREDIQT method (process and the prediction models) would you recommend?

12. Do you have further comments or suggestions?

APPENDIX 5: THE FEEDBACK RECEIVED THROUGH THE EVALUATION TEMPLATE

This section summarizes the feedback provided by five respondents on a pre-defined evaluation template (see Appendix 4). The summary is based on contents analysis of the answers obtained. All respondents are affiliated with the customer organization. Table II shows the background of the respondents, as reported through answers on **question 1** from the template.

TABLE III
RESULTS OF VALIDATION BASED ON 9 INDEPENDENT CHANGES

DV	**Estimated - old** (leaf node)	**Simulated - old** (root node)	**Simulated - Estimated** (root node)
1. DV dedicated to *Maintainability*	-0,01	0	0
1. DV dedicated to *Maintainability*	0,01	0	0
2. DV dedicated to *Maintainability*	-0,1	-0,01	0
2. DV dedicated to *Maintainability*	-0,02	0	0
2. DV dedicated to *Maintainability*	-0,16	-0,01	0,02
1. DV dedicated to *Usability w.r.t. contents*	0,5	0,01	-0,01
1. DV dedicated to *Usability w.r.t. contents*	-0,15	0	0,07
2. DV dedicated to *Usability w.r.t. contents*	-0,05	-0,01	0
2. DV dedicated to *Usability w.r.t. contents*	-0,07	-0,01	0,01

On **question 2**, the main strengths pointed out are: "The PREDIQT method is useful and it suits well the problem addressed"(R2), "Going for structure, going for reuse, utilizing group judgment and calculation of scores"(R1), "It was a way to in a systematic manner divide the problem in smaller parts, and then aggregate the quality level for the whole model"(R3), "Modeling concept – propagation of assessments"(R4). A weakness repeatedly pointed out is the missing formal mapping of the parameter estimates to the model, i.e. the parameter estimates may be too sensitive to the context and the interpretation (R1, R3, R4, R5). Complexity of the method and need for better tool support were pointed out by R5 and R1, respectively.

On **question 3**, all five respondents agreed that the models facilitate communication, knowledge exchange and understanding of the three aspects specified. R1 argues that "the workshops force people to communicate and harmonize into one model; the system is clarified and parts of architecture are disclosed and discussed; the most important part is assigning estimates on quality characteristics, which forces people to make statements". R1 however points out that the semantics of the characteristics should be more formal and the process of their harmonization in the group more strict. R2 argues that "the method provides a good model of the system, which can be communicated around; when a multi-disciplinary group manages to make a model of a complex problem and communicate around it, you have achieved a good result; when you additionally can make predictions based on the model, the result is even better." R3 argues that the communication is to a lesser degree efficient towards people outside the expert group, while R5 argues that the models are more useful for mapping consequences of changes, then for providing sufficiently precise predictions.

On **question 4**, R1 points out that the effort needed is reasonable from a typical management consulting perspective, but in an engineering context, more effort should be directed towards specific parts. R2 focuses on the benefit from useful discussions and insight into other's understanding of the system and the model. R3 thinks it is difficult to be a part of such a process without using more time, and expresses that more time should have been used on verification, while R4 and R5 think that the time needed for modeling is extensive.

On **question 5**, the answers vary. R2 and R3 point out that the process itself and the fact that the participants are encouraged to provide numbers is important for improved understanding. R2 thinks the uncertainty challenge lies in the estimates, while R1 sees the main uncertainty challenge in the Design Models, while Quality Models are in accordance with the goal. R3 expresses that the uncertainty of Design Models and Quality Models comes from unknown usage profile, while DVs give clear dependencies and propagation of assessments. R5 emphasizes that non-linearities are difficult to explicitly model in the Design Models and the DVs.

On **question 6**, R1 confirms the intention to use the models developed in the future, for purpose of architecture development and dependability analysis. R2 and R3 express the wish to use the method in future projects, given that financing can be provided. R4 intends to use the prediction models if they can be tailored to specific use cases, while R5 writes "I believe the model can be used to understand and predict the result/risk in different changes".

On **question 7**, R1 expresses "it has already served the purpose in creating understanding and analysis. If incorporated with more tool support, I think it can be utilized in practice". R2 expresses that PREDIQT is very much better than no method, but it is unknown what it takes for it to be perfect. R3 and R4 express that the benefit from the method and quality of the predictions depend on the modeling skills and granularity of the models. R5 points out the challenge of interpreting the predictions due to the lack of documentation of the assumptions made during the parameter estimation.

On **question 8**, R2 expresses two main challenges "access to competent resources to make the models and interpretation of the predictions and the corresponding uncertainty which requires competence". R3 points out three aspects: "be sure that you have modeled the most important aspects; models need to be verified; define the values in a consistent way". R4 sees the uncertainty challenge in the fact that the changes are marginal and therefore give small effects on the numbers, while R5 relates uncertainty to the insufficiently formal interpretation of the parameter values due to the assumptions made during their estimation.

On **question 9**, R2 expresses that the method "reduces uncertainty at change, but does not eliminate it; but it does systematize the uncertainty and reduce it sufficiently so that the method absolutely is valuable". R3 sees the discussion of the quality characteristics and agreement upon the most important characteristics, as the main benefit. R4 answers "the

final model", while R5 writes: "the Design Model itself and DVs; making the model forces you to consider all parts and their dependencies".

On **question 10**, R1 emphasizes the harmonized system understanding. R2 expresses that the totality is important, and it is difficult to select certain models that are more important, just as is the case when thinking if wheels or the motor are most important on a car. R3 answers: "since I think the process and not the prediction is most useful, the model is just a tool to facilitate the discussion". R4 and R5 answer "DVs", and R5 also adds: "making the DVs forces you to consider all parts and their dependencies".

On **question 11**, R1 emphasizes tool support, stricter workshops, and in-advance preparation of the experts. R2 also points out simpler tool support, which would enable large-scale use of the method by many end-users. Furthermore, R2 points out the need for increased traceability between the models, so that the relationships (and impacts among the models) are more explicit to the user, instead of model consistency/validity keeping (after a change deployment) being a manual task. R2 also expresses the challenge of keeping the models consistent with the underlying system which is undergoing an evolution and needs continuous support for prediction of the impacts of changes. Furthermore, R2 asks if existing system models can be reused and serve as the Design Models, as it may enable use of existing methods and analysis related to domain specific notations. R3 expresses that "it was difficult to define a quality value. Maybe the method should recommend how the teams could work to get consistent values through the model". R4 suggests "detailed views of parts of the system", while R5 again expresses the uncertainty regarding the possibility of modeling the system realistically in a linear model.

No respondent had comments in relation to **question 12**.

APPENDIX 6: THREATS TO VALIDITY AND RELIABILITY

The validity of the findings with respect to 1) the performance of the PREDIQT method; and (2) the results of the evaluation of the predictions based on the thought experiment and the overall assessment, depends to a large extent on how well the threats have been handled. In addition to reliability threats, four types of validity threats, presented in [5, 20], are addressed: construct validity, conclusion validity, internal validity and external validity.

Reliability is concerned with demonstrating that the operations of the case study can be repeated with the same results. Of course, an industrial case like this can never give solid repeatable evidence. There are too many contextual factors influencing what happens, as has been pointed out in Section VI with respect to assumptions and interpretations related to parameters. In the case of our analysis it may be argued that we should have focused more on verification, used historical data and performed more measurements rather than using merely expert judgments, but such data were not available. Instead, various excel sheets and other relevant documents (that models and numbers could be mapped to) were extensively used during Design Model development and DV estimation. We did however perform thought experiments on known and implemented changes (whose effects on quality are well known) at several meetings, on order to validate the models and evaluate the predictions. Still, the thought experiment based evaluation does have more weaknesses compared to the measurement based one.

The active involvement of the analyst may to some degree have influenced the analysis and needs to be reflected upon. However, since the role of the analyst is included in the method, the analyst should not have been a significant source of bias. The involvement of the analyst has also allowed better insight into the process and a more detailed evaluation of the results.

Construct validity concerns whether we measure what we believe we measure. Both the model development and the thought experiment relied on the subjective estimates provided by domain experts. There was some turnover among the domain experts, but two of them participated throughout the case study. The simulations themselves were conducted by the method developer after and independently from the thought experiment.

The change specifications included diverse, non-overlapping changes covering major parts of the prediction models. The quality attribute specific DVs were very complex, which minimizes the possibility that the domain experts were able to remember the DVs and thus quickly calculate propagation of the changes during the thought experiment. The variance of the discrepancy between the simulated and the thought experiment based values is quite high. All this considered, the risk that the prediction models, the change impact simulations and the thought experiment based estimates were consistently wrong, should be relatively small. Hard evidence in the form of measurements to validate the correctness of the predictions would have been desirable, but this was unfortunately impossible within the frame of this case study.

Another threat to the construct validity is the possible discrepancy in the understanding of the prediction models, particularly the Quality Models, by the domain experts. In case different assumptions have been made implicitly, they are not documented in the estimates. However, the active involvement of the participants in the model development and the ability of reaching agreement during development, validation and use of the models, do not give reason to assume that the models express else than specified.

Conclusion validity concerns the composition of participants and the statistical analysis. Statistical power was limited, due to low number of participants. The careful selection of experienced participants and the variety of the changes should have compensated for some of this.

Internal validity concerns matters that may affect the causality of an independent variable, without the knowledge of the researcher. Causality is present in decomposition of the quality characteristics to quality indicators, in the development of the DVs, and in the selection of the relevant indicators during the estimation of the DV parameters. We

have addressed these threats by involving the domain expert panel in all model development and validation, by actively using the available evidence and documentation during the development of the Design Models and the estimation, and by using the established ISO 9126 standard for the decomposition and definition of the quality notions.

Additionally, in order to ensure that the DV structure fulfills the requirements regarding completeness, orthogonality and a dependency model developed from the perspective of the quality characteristic in question, the analyst has systematically guided the domain experts and explained the principles, during the DV structure development. The Quality Models and the Design Models were also actively used during the DV development, as they provide the underlying quality definitions and system representation, respectively.

External validity concerns the generalization of findings of this case study to other contexts and environments. The target system is representative for the systems intended to be within the scope of the PREDIQT method. Moreover, this is the second trial of PREDIQT in a real-life setting and both trials have given promising results and useful insights. Both trials have given strong indications of feasibility of the method, reported similar benefits (understanding of system architecture and its quality, usefulness of estimates particularly when interpreted relative to each other, and usefulness of the process) and undergone the same stages of the PREDIQT process. More measurements and their analysis were however provided during the verification in the first case study (while thought experiments were performed during the evaluation). This case study has involved more domain experts and both the validation and the evaluation have merely been based on thought experiments. The individual DVs developed in each case study were of similar size. The first case study resulted however in three DVs – one for each quality characteristic, while this one resulted in four DVs – two for each quality characteristic. Apart from that, there is no significant difference in the size or complexity of the prediction models between the two case studies.

No particular customizations of the method were needed for this trial. Thus, we have reason to believe that it should be possible to reapply PREDIQT in another context.

Appendix 7: Related work

The PREDIQT method for model-based prediction of impact or architectural design changes on system quality characteristics makes use of models that capture the system design, the system quality notions and the interplay between system architecture and quality characteristics, respectively. The predictions result in propagation paths and the modified values of the parameters which express the quality characteristic fulfillment at the different abstraction levels. The PREDIQT method aims at establishing the right balance between the practical usability of the models through the simplicity of the model structures, and the soundness of the predictions through a multi-stage structured process.

We are not aware of other approaches that combine these elements in this way. However, the issues of metrics estimation, system quality and the various notations for modeling system architecture, have received much attention in the literature.

According to [8], most prediction models use size and complexity metrics to predict defects. Others are based on testing data, the quality of the development process, or take a multivariate approach. In many cases, there are fundamental statistical and data quality problems that undermine model validity. In fact, many prediction models tend to model only part of the underlying problem and seriously misspecify it.

Bayesian networks (BNs) [15, 11] allow incorporating both model uncertainty and parameter uncertainty. A BN is a directed acyclic graph in which each node has an associated probability distribution. Observation of known variables (nodes) allows inferring the probability of others, using probability calculus and Bayes theorem throughout the model (propagation). BNs are however demanding to parametrize and interpret the parameters of. This issue has been addressed by [18] where an analytical method for transforming the DVs to Bayesian networks is presented. It also shows that DVs are, although easier to relate to in practice, compatible with BNs.

PREDIQT is compatible with the established software quality standard [1], which is applied in this case study. The goal/question/metric paradigm [3, 2] is a significant contribution to quality control which also is compatible with PREDIQT and can be used for development of Quality Models and design of a measurement plan [9, 7].

[6] and [14] provide surveys of the software architecture analysis methods (SAAM, ATAM, ALPSM, SAEM etc.). Compared to PREDIQT, they are more extensive and provide a more high-level based architecture assessment of mainly single quality characteristic (maintainability or flexibility). Furthermore, they are not predictive, do not incorporate measurement, and quality is defined and quantified differently. ATAM [13, 12, 4] is, for example, more coarse-grained than PREDIQT in terms of both quality definitions and measurement. PREDIQT allows a more fine grained analysis of several quality characteristic and their trade-offs simultaneously and with limited effort. Hence, an integration of the two may be worthwhile examining.