SCORING A SOFTWARE DEVELOPMENT ORGANIZATION WITH A SINGLE NUMBER

BY

RYAN M. SWANSTROM

A dissertation submitted in partial fulfilment of the requirements for the

Doctor of Philosophy

Major in Computational Science and Statistics

South Dakota State University

2015

SCORING A SOFTWARE DEVELOPMENT ORGANIZATION

WITH A SINGLE NUMBER

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy in Computational Science and Statistics degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidates are necessarily the conclusions of the major department.

Gary Hatfield, Ph.D.
Dissertation Advisor Date

Kurt Cogswell, Ph.D.
Head, Department of Mathematics and Statistics Date

Dean, Graduate School Date

"What you have here is small data."

Clay Campbell

"Every day, three times per second, we produce the equivalent of the amount of data that the Library of Congress has in its entire print collection, right? But most of it is like cat videos on YouTube or 13-year-olds exchanging text messages about the next Twilight movie."

Nate Silver

ACKNOWLEDGEMENTS

I would like to acknowledge the generous support I received from my family. Thank you to Emily for providing encouragement and the final push to get me to eventually finish. Thank you to Ainsley, Porter, Trey, and Ryker for always providing a smile.

Next, I would like to thank my advisor, Dr. Gary Hatfield, for the encouragement and guidance along the way. Also, thank you to Clay, Jess, and Chris for helping me formulate the initial idea.

Finally, I would like to thank my Mom for providing a final grammatical proof-reading of the dissertation.

CONTENTS

ABBREVIATIONS

CDF	Cumulative Distribution Function
CEO	Chief Executive Officer
CMMI	Capability Maturity Model Integration
CRI	Cumulative Result Indicator
IEEE	Institute of Electrical and Electronics Engineers
KPI	Key Performance Indicator
KRI	Key Result Indicator
NOM	Number of Methods
PI	Performance Indicator
PROD	Production
RI	Result Indicator
SDLC	Software Development Life Cycle
SDLC-AE	SDLC Analytic Engine
SDO	Software Development Organization
SEMAT	Software Engineering Method and Theory
SIT	Systems Integration Testing
SLOC	Source Lines Of Code
SQL	Structured Query Language
UAT	User Acceptance Testing

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

SCORING A SOFTWARE DEVELOPMENT ORGANIZATION WITH A SINGLE NUMBER

RYAN M. SWANSTROM

2015

Nearly every large organization on Earth is involved in software development at some level. Some organizations specialize in software development while other organizations only participate in software development out of necessity. In both cases, the performance of the software development matters. Organizations collect vast amounts of data relating to software development. What do the organizations do with that data? That is the problem. Many organizations fail to do anything meaningful with the data.

Another problem is knowing what data to collect. There are many options, but certain data is more important than others. What data should a software development organization collect?

This paper plans to answer that question and present a framework to gather the right information and provide a score for an organization that produces software. The score is not to be comparative between organizations, but to be comparative for a specific organization over time.

The primary goal of this work is to provide a general framework for what a software development organization should measure and how to report on those measurements. The focus is providing a single number to represent the entire organization and not just the development efforts. That single number is considered the Cumulative Result Indicator (CRI) score. The secondary goal of this work is to provide a framework for storing the necessary data.

1 INTRODUCTION

Software is becoming a vital part of companies. In 2011 Marc Andreessen, co-founder of the venture capital firm Andreessen-Horowitz, famously claimed, “Software is Eating The World” [1]. His argument was for the ever increasing importance of software in all organizations big and small regardless of the industry. With this important declaration, the production of new software is going to be critical. Just as important will be the effective measurement of how this software is produced.

This dissertation provides a technique for a Software Development Organization (SDO) to create a single number score which indicates the overall performance of the organization. The score is based upon data collected for five key result indicators of an SDO: quality, availability, satisfaction, schedule, and requirements. It is not meant to be comparative between organizations, but to form a historical baseline for a specific organization. The single number is targeted for upper-level management who need a quick and simple strategy to evaluate the performance of the organization.

1.1 OVERVIEW

An SDO is no different than any other business or organization. There are: tasks to be completed, goals to achieve (or miss), and measurements to be analyzed. One difficulty with software development is the varied number and amount of measurements to be used. It can be difficult to determine the correct activities to measure and the appropriate mechanism to report the measure. This has led organizations to either collect too little information or to collect too much information. Another problem is the inconsistency of the reported measures. It is difficult to compare historical performance if the same measurements are not consistent throughout the recent history of an organization.

SDOs need a framework to define what measurements should be tracked and how those measurements should be reported. The Cumulative Result Indicator (CRI)

framework provides a solid foundation for a consistent evaluation. CRI analyzes the historical performance of an SDO to create a baseline in order to provide a broad view of the overall organization. It is common for software development organizations to measure and focus solely on the source code being produced. However, an SDO does more than just produce source code. There is documentation to be written, testcases to be created, systems to be deployed, and decisions to be made. The framework provides an evaluation of the overall SDO, not just the source code.

The framework will produce a single number score for each of the five result indicators as well as a single overall score. It will be able to provide a quick evaluation of the organization. The scores will enable performance to be consistently measured and compared.

Other attempts at evaluating an SDO have been presented, but none produce a single number score for the entire effort of the SDO. The following are attempts to evaluate all or parts of software development.

1.2 TERMINOLOGY

Like any other business domain, the software engineering field has a number of specific terms. Many of these terms will be used throughout the remainder of the document, so definitions are provided.

Application - A software system or a collection of other applications

Release - A collection of projects being put in production on a specified date

Project - A body of work involving zero or more applications in preparation for a release

SIT (Systems Integration Testing) - The initial step of testing after the development phase of the SDLC. This is typically performed by members of the SDO. It is validation that all the software components function together as expected.

UAT (User Acceptance Testing) - The final step of testing when a select few members of the user group are invited to validate the software system. Once validation has occurred for UAT, the software system is ready to proceed to production

PROD (Production) - The software has been released to the final audience.

Defect - "A software defect is a bug or error that causes software to either stop operating or to produce invalid or unacceptable results" as quoted from Capers Jones [45]. It is important to mention that even though defects are typically found in the computer code, a defect should not be isolated to just code. A poorly written requirement or missed testcases can both be considered a defect. Other common names for a defect are: bug, error, fault, or ticket.

1.3 CMMI

The Capability Maturity Model Integration (CMMI) is one of the most widely acknowledged models for process improvement in software development. CMMI offers a generic guideline and appraisal program for process improvement. It was created and is administered by the Software Engineering Institute at Carnegie Mellon University [16]. While the CMMI is not specific to software development, it is often applied in software development settings. CMMI certification is required for many United States Government and Department of Defense contracts.

CMMI-Dev is a modification of the CMMI specific to the development activities applied to products and services. The practices covered in CMMI-Dev include project management, systems engineering, hardware engineering, process management, software engineering, and other maintenance processes. Five maturity levels are specified, and they include the existence of a number of process areas. The five maturity levels and the process areas are specified as follows.

CMMI MATURITY LEVEL 1 - INITIAL A maturity level 1 organization consists of an impromptu and chaotic process. While working products are still produced, the results are often over budget and behind schedule. A level 1 organization will also have difficulties repeating a process with the same degree of success. These organizations typically rely on the heroic efforts of certain individuals.

CMMI MATURITY LEVEL 2 - MANAGED A maturity level 2 organization has a policy for planning and executing processes. The processes are controlled, monitored, reviewed, and enforced. The practices are even maintained in times of stress. The following process areas should be present at maturity level 2.

- Configuration Management (CM)
- Measurement and Analysis (MA)
- Project Monitoring and Control (PMC)
- Project Planning (PP)
- Process and Product Quality Assurance (PPQA)
- Requirements Management (REQM)
- Supplier Agreement Management (SAM)

CMMI MATURITY LEVEL 3 - DEFINED A maturity level 3 organization has well-understood processes that are described in standards, tools, procedures, and methods. The organization has standard processes that are reviewed and improved over time. The major differentiators between level 2 and level 3 is the existence of standards and process descriptions. A level 2 organization will have processes that are inconsistent across projects. A level 3 organization will tailor a standard process for each project. Also, level 3 processes are described with much

more rigor. In addition to the process areas found in level 2, the following process areas should be present at maturity level 3.

- Decision Analysis and Resolution (DAR)
- Integrated Project Management (IPM)
- Organizational Process Definition (OPD)
- Organizational Process Focus (OPF)
- Organizational Training (OT)
- Product Integration (PI)
- Requirements Development (RD)
- Risk Management (RSKM)
- Technical Solution (TS)
- Validation (VAL)
- Verification (VER)

CMMI MATURITY LEVEL 4 - QUANTITATIVELY MANAGED

A maturity level 4 organization has quantitative measures for quality and process performance. The measures are based upon customer needs, end users, and process implementers. The quality and process performance are understood mathematically and managed throughout the life of a project. Level 4 is characterized by the predictability of the process performance. In addition to the process areas found in level 2 and 3, the following additional process areas should be present at maturity level 4.

- Organizational Process Performance (OPP)
- Quantitative Project Management (QPM)

CMMI MATURITY LEVEL 5 - OPTIMIZING The final and pinnacle level of CMMI maturity is level 5. A maturity level 5 organization continually improves processes based upon quantitative measures. The major distinction from level 4 is the constant focus on improving and managing organizational performance. A maturity level 5 organization has well-documented standard processes that are tracked and enforced as well as a focus on continual improvement of the processes based upon quantitative measures. In addition to the process areas of the previous maturity levels, maturity level 5 should contain the following process areas.

- Causal Analysis and Resolution (CAR)
- Organizational Performance Management (OPM)

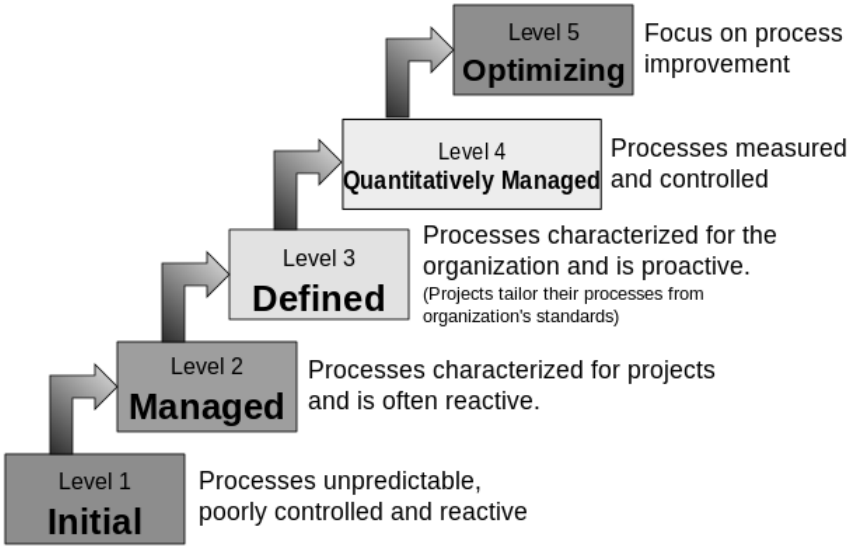

Figure 1: Characteristics of CMMI, image adapted from [30]

A visual description of the CMMI maturity levels can be seen in Figure 1. While CMMI-Dev does provide an excellent framework for improving a process, it is entirely focused on process improvement. It does not provide guidelines for evaluating the final product. Also, it does not provide a specify mechanism for evaluating or scoring the

progression through the maturity levels. An indicator is still needed to quantify the overall performance of an organization, not just the compliance to standard processes.

1.4 PREVIOUS SOFTWARE EVALUATION WORK

An example of scoring software development is presented by Jones [46]. The methodology looks for the presence of various techniques used in software engineering. The methodology provides a score based upon the productivity and quality increase of the technique being evaluated. Points are positive or negative based upon the presence of various techniques. Two examples of such techniques are: automated source code analysis and continuous integration. The end result is a score in range $[-10, 10]$. While the result is a single number score, it does not account for the entirety of the software development life cycle.

Constructive Cost Model (COCOMO) is a software cost estimation model created by Boehm [8]. It combines future project characteristics with historical project data to create a regression model to estimate the cost of a software project. The original version developed in 1981 was focused on mainframe and batch processing. An updated version, named COCOMO II, was created by Boehm in 1995 to be more flexible for newer development practices such as desktop development, off the shelf components, and code reuse. COCOMO provides a nice algorithm for making decisions regarding building or buying software products. It does not provide an algorithm to review and modify past performance based upon estimates. COCOMO II can be a useful tool for estimating the time and costs within an SDO, but it only provides an estimate and not an evaluation of the actual performance.

Sextant is a visualization tool for Java source code [90]. Sextant provides a graphical representation of the information related to a software system. The tool provides the capability to provide custom rules which are specific to the domain or application. However, Sextant only provides metrics and analysis of the software code. It provides no

information regarding the rest of the software development life cycle. Also, the primary output of Sextant is visual graphs. While these graphs do provide useful information, they do not provide a single number to determine the performance of the software.

Another promising research area is *process mining* [87]. As stated by Wil van der Aalst [86], "The idea of process mining is to discover, monitor and improve real processes (i.e., not assumed processes) by extracting knowledge from event logs readily available in today's systems." Creating software involves many processes. Numerous logs of raw data are collected. One application of process mining in the area of software development was an algorithm and information for measuring a software engineering process from the Software Configuration Management (SCM) [73]. The technique creates process models to understand the process of developing software and code. It is less focused on the output and results, but it is more focused on adherence to a specified process.

Process mining has also been applied to decision making regarding software upgrades [88]. Historic and current logs can be processed and evaluated. Then small pilot groups can be offered the upgrade, and the new logs will be processed and evaluated. Thus, process mining can be beneficial for new software implementation. More research needs to be done applying process mining to the other processes involved with software development, such as other documentation, testing, and implementation.

Although process mining can be useful for analyzing software development data, it has not yet been applied to the entirety of an SDO. It also does not focus on the results of the organization. It focuses on process conformance, which can be very important, so it is limited in the ability to evaluate the entire organization.

Much work has been done to determine metrics for source code, in fact entire books have been written on the topic of software metrics [44], [68]. Yet, organizations still struggle to measure the production of software. Little work exists for scoring the entire SDO.

1.4.1 SEMAT

Software Engineering Method and Theory (SEMAT) is claiming to be the "new software engineering" [41]. The authors rightfully claim that software engineering lacks an underlying theoretical foundation found in other engineering disciplines. This lack of theory has led software engineering to not be engineering, but rather a craft. The goal of SEMAT is to merge the craftsmanship and engineering to provide a foundation for software engineering. The primary initiative of SEMAT has been the creation of a kernel for software engineering. The kernel is the minimal set of things common to all software development endeavours. The three parts to the kernel are:

1. **Measurement** - There must exist a means to determine the health and progress of an endeavour.
2. **Categorization** - The activities must progress through categories during an endeavour.
3. **Competencies** - Specific competencies will be required for completing activities.

The kernel defines alphas, which are seven dimensions with specific states for measuring progress. The seven dimensions are:

1. Opportunity
2. Stakeholders
3. Requirements
4. Software Systems
5. Work
6. Team
7. Way of Working

Although SEMAT is very promising, the development is not yet complete. Adoption is limited so the technique has not been validated on many actual software engineering endeavours. Although SEMAT does include a part for measurement of progress, it does not specify how the measurement is to be performed.

1.4.2 SOFTWARE QUALITY

Software quality is one of the most well studied aspects of software development. Most of the work focuses on either the problems with the software or the source code. Quality and the number of problems with the software are inversely related, more problems means lower quality. Quality is easy to measure, but that measurement is usually very software specific. It is easy to find that some software has X number of problems, but it is nearly impossible to determine whether the quality of that software is better or worse than some other software with X defects. One piece of software can be larger[1] or more complex. Thus, finding a value for quality is easier than interpreting that value. No matter the interpretation, the goal is to decrease the number of problems with the software. Top 10 lists have been created for techniques to remove problems from software [11]. A number of different techniques or best practices for preventing defects have been proposed [27]. These are all strategies to identify or remove the problems before the software is completed and released to users.

Another aspect of software quality is the complexity of the source code. More complex code results in more maintenance efforts and more chances for problems. Some numerical measures for the complexity of source code have been created. The most common examples are McCabe [60] and Halstead [33]. However, the measures on source code only explain part of the software development life cycle.

Another measurement of quality can be the cost per defect, also known as the cost

[1]Saying a piece of software is larger can be a rather arbitrary statement. It can mean the software requires more computing time, has more lines of code, more documentation, more hours spent on development, or some other arbitrary measure.

to fix a problem. As seen in [48], this measurement has problems because the lowest cost per defect will occur on software with the most problems. Therefore, the lowest cost per defect is actually the lowest quality as well. Due to this difficulty and others, a number of other models have been created for evaluating the quality of software [63]. While all of the models have merit in certain situations, the measures of quality must be combined with other measures in order to provide an overall evaluation of an SDO.

1.4.3 SOFTWARE ANALYTICS

One area of research that is focusing on the evaluation of SDOs is *software analytics*. Software analytics is less focused on evaluation and more so on all sorts of analysis of software data. The earliest variants of software analytics were disguised as applications of data mining techniques to software engineering data in the late 1990s [25], [31], [74]. Later the field began to emerge more heavily, but still remained primarily methods of data mining applied to software engineering [32], [51], [83], [91]. The term software intelligence was proposed for the field of study [35], but eventually the term software analytics became the dominant term for referring to the field of study [13], [92].

The goal of software analytics is to extract insights from software artifacts to aide practitioners in making better decisions regarding software development [93]. The three main focus areas of software analytics are:

1. **User Experience** - How can the software enable the user to more easily or quickly accomplish the task at hand?

2. **Quality** - How can the number of problems with the software be decreased?

3. **Development Productivity** - How can the processes or tools be modified to increase the rate at which software is produced?

Later, Martin Shepperd in [34] identified three important questions that software analytics must address:

1. “How much better is my model performing than a naive strategy, such as guessing [. . .]?”

2. “How practically significant are the results?”

3. “How sensitive are the results to small changes in one or more of the inputs?”

These are three important questions that should be addressed when presenting any results in software analytics. The research needs to demonstrate clear advantages for practitioners. The work presented in this dissertation will address both the three main focus areas and the three important questions of software analytics.

Lavazza, Morasca, Taibi, and Tosi focus specifically on the source code; analyzing the complexity, size, and coupling [84]. They created a theoretical framework for dynamic measurements instead of traditional static measures. Letier and Fitzgerald discuss how to choose the correct tools and techniques to analyze software data [57]. A goal model is produced that matches the data analysis methods with the goals of the software stakeholders. The method does not focus specifically on analysis of the development of software.

Software Development Analytics is a subfield of software analytics [62]. It focuses specifically on the analytics of the development of software, however not the overall performance of the software. Hassan points out in [34] that software analytics needs to go beyond just the developers. Everyone and everything involved in the development of software produces some data and that data can be meaningful. The insights from non-developer data has the potential to yield important results as well. Software development produces many valuable pieces of datum that can be analyzed [59]. Just a few of the pieces of datum are: email communication, bugs, fixes, source code, version control system histories, process information, and test data. Examples of this type of data can be found in the PROMISE Data Repository [61].

All of these techniques are of no use if the correct data is not available. Therefore,

it is important to identify the information that is needed to properly perform software development analytics. Unfortunately, there are vast amounts of information that need to be collected to meet the analytic needs of developers and managers [14]. When the data and tools exist, the analytics should help an organization with the following tasks.

- Evaluate a project
- Determine what is and is not working
- Manage risk
- Anticipate changes
- Evaluate prior decisions

In order to store the data, appropriate tools are needed. Microsoft is working on developing some tools for the analysis of the development of software [19], [93]. Microsoft has developed StackMine, a postmortem tool for performance debugging, and CODEMINE, a tool for collecting and analyzing software development process data. Both tools provide analytical insight for various aspects of the software development process, however neither tool covers all aspects of software development. These tools are currently early in development and the adoption of the tools by practitioners is still unknown. One of the reasons for the slow adoption of new tools is the inherent difficulty of producing new tools for the software development process [80]. A tool that works fantastic for one team might not automatically apply to another team. The people creating the tools need to be acutely aware of the needs of the technical practitioners that will be using the tools.

Iqbal, Ureche, Hausenblas, and Tummarello introduced a methodology named Linked Data Driven Software Development (LD2SD) which is a collection of various software artifacts into linked data [40]. This is one of the original attempts to collect software engineering data. The methodology links version control, discussion forums, and

issue tracking data. The result is web-scale integration of data, but the actual benefits are still uncertain.

After the proper tools are in place to collect the necessary data on software development and software analytics are being properly implemented, an obvious next step is the application of gamification to software development. Gamification is "the process of making activities more game-like" [89]. Some of the benefits of gamification are higher productivity, added competition, and greater enjoyment. Prior attempts at gamification of software development focus only on the computer programming phase [78]. Others focus on defining a framework for gamification within the software development process [42]. There are even some indications that gamification might help increase software quality [24]. While this dissertation will not focus on gamification, it is important to note that an implementation of an evaluation technique for software development could be implemented simultaneously with a gamification strategy. Both will require new collections of data and new reporting.

Overall, there exist many attempts to evaluate portions of an SDO. None of the attempts provide a single number score for the entirety of the organization. Most of the techniques focus on specific portions of the software development life cycle, namely the development portion. Plus, there are many tools that need to be created for software analytics to provide all the value that it promises.

1.5 ORGANIZATION OF THE WORK

The remainder of this dissertation is divided into seven chapters. Chapter 2 provides an overview of software, software development life cycles, software engineering, and software development organizations. Chapter 3 introduces some existing techniques for measuring an SDO. Chapter 4 then provides an explanation of the Cumulative Result Indicator (CRI). It will present the essential elements for calculating the CRI, as well as the formulas, framework, and data necessary to produce the CRI. Chapter 5 provides a

technological framework for storing the current and historical CRI values. Chapter 6 demonstrates how CRI can be implemented in a software development portion of a large financial institution. Chapter 7 discusses some possible future directions for further enhancements. Chapter 8 concludes the dissertation with a summary of the results.

2 A SOFTWARE DEVELOPMENT ORGANIZATION

A *Software Development Organization (SDO)* is any organization or subset of an organization that is responsible for the creation, deployment, and maintenance of software. Many times an SDO is a company that produces software. Other times, an SDO is contained within the Information Technology department of a larger organization. Some of the job roles with an SDO are: software engineer, system administrator, software quality analyst, programmer, database administrator, and documentation specialist.

2.1 WHAT IS SOFTWARE?

Numerous definitions can be found for the term *software*. Software is more than just computer programs. According to Ian Sommerville [79], "Software is not just the programs but also all associated documentation and configuration data which is needed to make these programs operate correctly." This is the definition used for the remainder of this dissertation.

2.2 THE SOFTWARE DEVELOPMENT LIFE CYCLE

The discipline of software engineering has created a workflow for developing software. This workflow is called the *Software Development Life Cycle (SDLC)* . SDLC can be defined as [75]:

> [...] a conceptual framework or process that considers the structure of the stages involved in the development of an application from its initial feasibility study through to its deployment in the field and maintenance.

While the SDLC states what needs to be done, there are numerous models that formalize exactly how to perform the SDLC. The models contain steps that are commonly referred to as a phases. A few of the popular models are described below.

2.2.1 WATERFALL

The waterfall model is the oldest and most influential of the SDLC models. It was first presented at a Navy Mathematical Computing Advisory Panel in 1956 by Herb Benington [7]. Figure 2 shows the model Benington outlined for producing large software systems. In 1970, Benington's model was modified by Royce [72]. Royce produced an updated version of the diagram seen in Figure 3 which provides some loops to go back to a previous phase in the workflow.

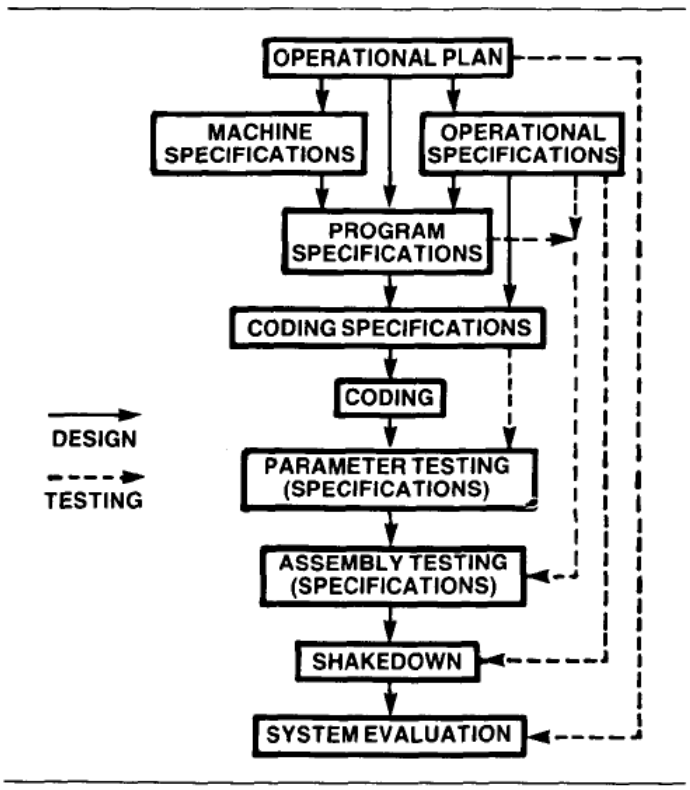

Figure 2: Benington's original diagram for producing large software systems, adapted from [7]

The modern version of the waterfall model specifies that each phase needs to be

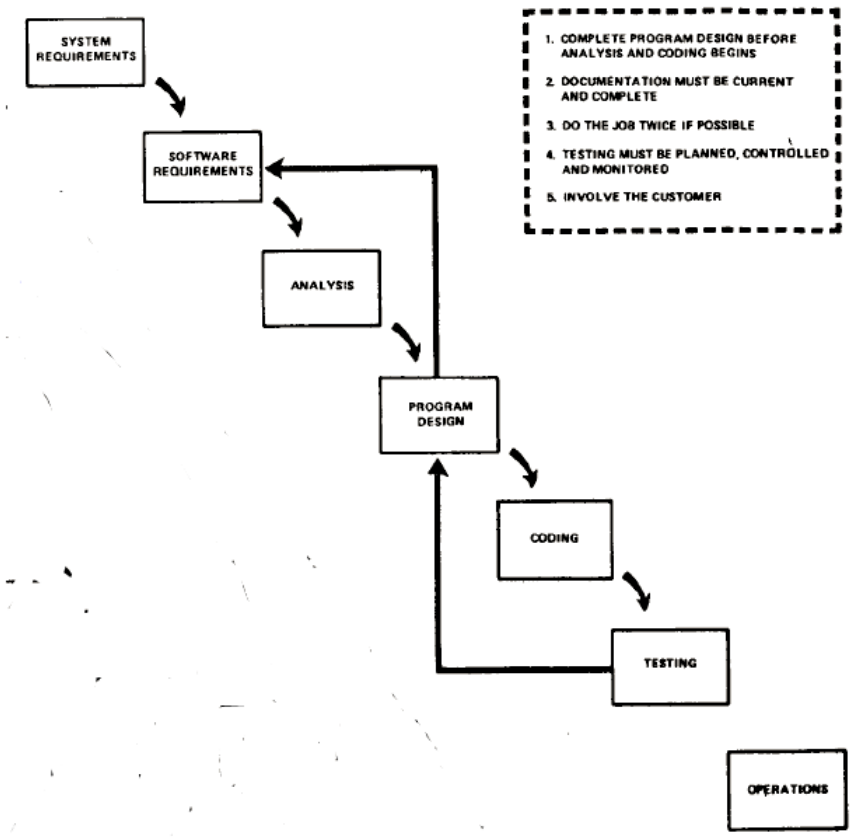

Figure 3: Royce's version of the waterfall model for producing software systems, adapted from [72]

entirely completed before moving onto the next phase. Some small amount of overlap is permitted and looping occurs but both actions are discouraged and should be limited. A modern diagram of the waterfall model can be seen in Figure 4.

Waterfall has some excellent features such as: simple to understand, easy to plan, and well-defined phases. However, waterfall lacks the flexibility required of many software systems built today [58]. Due to the fact the phases are so sequential, it makes changes during the life cycle difficult and expensive if not impossible. Therefore, other models of SDLC have been created to address the lack of flexibility of the waterfall model. Notice, the other models are adaptations of waterfall.

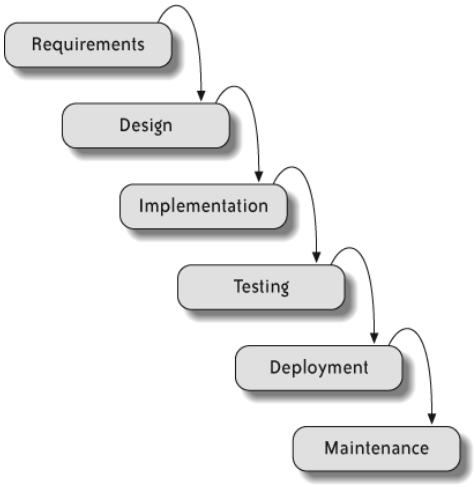

Figure 4: Modern Waterfall Diagram, adapted from [36]

2.2.2 SPIRAL

The spiral model for software development was presented by Boehm in 1986 [9], [10]. The goal of the spiral model of software development is very risk-driven. A software project will start with many small and quick iterations. Each iteration will cover the following 4 basic steps.

1. Determine Objectives
2. Identify Risks
3. Develop and Test
4. Plan Next Iteration

This model allows software to be built over a series of iterations without risking too much time or effort in any single iteration. Spiral requires a very adaptive management approach as well as flexibility of the key stakeholders [75]. It can also be difficult to identify risks that will occur in future iterations. Figure 5 provides a bit more detail on the iterations and the overall process.

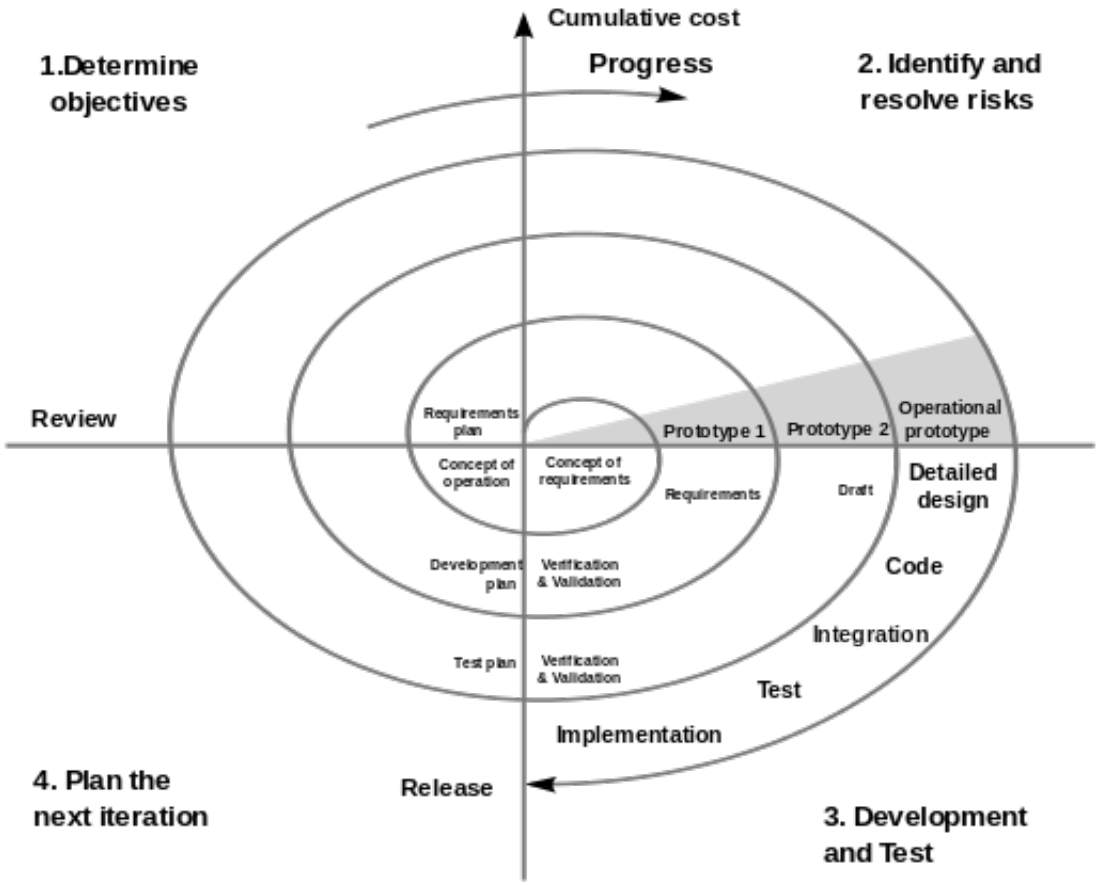

Figure 5: Spiral SDLC Model [12]

2.2.3 AGILE

Agile software development has arisen due to the inability of the waterfall and other models to adjust to changes during the development cycle. Agile software development is a group of SDLC models that operate under the influence of the following four key principles [6].

1. Individuals and interactions over processes and tools
2. Working software over comprehensive documentation
3. Customer collaboration over contract negotiation
4. Responding to change over following a plan

Agile does not specify an implementation, but some specific models of agile SDLC are: eXtreme Programming, Scrum, Lean, Kanban and others [36], [64]. Agile models are very popular in many of today's software development organizations because the models work well for dynamic quickly changing applications such as web-based applications. Startups have largely adopted the Lean methodology for its ability to identify a minimum viable product[2] and reduce the time to market [29].

2.2.4 SDLC COMMONALITIES

Even with the large number of SDLC models currently being used by different SDOs, many commonalities exist among the models. The commonalities can be tied back to the steps of waterfall. All of the models exhibit, to some degree, the following phases. The only major difference is the scope, size, and duration of each phase. For example, the spiral model spends less time in each phase. The agile models produce less documentation and focus more on the implementation phase. Here are the five common phases in nearly all SDLC models:

1. **Requirements**

   The first phase is involved with defining what the software must do. Each piece of functionality is considered a requirement.

2. **Design**

   Before writing any code, the necessary infrastructure and involved software systems must be identified. This phase can serve as a guideline for the remaining phases. If done properly, this phase can greatly help the later phases.

3. **Implementation**

   Often the only phase of the SDLC that is measured, this is the phase where the actual computer code is written.

[2] A *minimum viable product* is a reduced version of software that contains only the bare minimum functionality required to meet the requirements.

4. Testing

This phase validates the expected functionality. Also, testing attempts to discover unexpected side affects of the software.

5. Deployment and Maintenance

All software must be correctly deployed and maintained. This phase is the most expensive and lengthy phase of the software development life cycle.

These five steps cannot cover everything that needs to be accomplished during a project. They just provide of rough guideline of what needs to be completed in order to ensure a more successful software release. Appendix A contains a more detailed list of the tasks necessary to complete the software life cycle.

2.3 WHAT IS SOFTWARE ENGINEERING?

Software Engineering as a term dates back to the 1968 North Atlantic Treaty Organization (NATO) conference [65], [85]. Over the years many definitions have been provided. Institute of Electrical and Electronics Engineers (IEEE) provides a definition that encompasses many of the other definitions. IEEE ISO610.12 defines software engineering as, "The application of a systematic, disciplined, quantifiable approach to the development, operation, and maintenance of software" [39].

Software engineering has struggled to determine the correct projects to complete [22]. Software projects are commonly behind schedule and over budget [15], [38], [49], [56] with nearly 20% of projects in the United States still failing [26]. As of 2015, Software engineering is still awaiting the professional status of more established fields such as medicine, law, and general engineering [47]. Organizations need a better technique to understand the past performance so they can better predict the future performance. Proper measurement will be essential to solidifying software engineering as certified profession.

3 MEASURING AN SDO

Anything can be measured [37]. Thus, an SDO can be measured. Proper measurement is crucial for improvement because without a starting point it is impossible to determine progress. Also, consistent reporting is essential for tracking historical performance.

The SDLC, like any process, needs to be properly measured. In order to accomplish proper measurement, three activities need to occur [28].

1. Identify Process Issues
2. Select And Define Measures
3. Integrate with the Software Process

This dissertation will focus on steps 1 and 2. The process issue is the overall effectiveness of the SDO. Section 4 will cover step 2 as it relates to an SDO. Step 3 will be different for each SDO, but Section 5 provides a bit of guidance for storing the correct information. It is up to the specific SDO to determine how and when the information is being stored.

Many methods have been used in the past to measure and evaluate SDOs. Some of the common methods will be explained in the following sections.

3.1 METRICS

A *metric* can be defined as a means of telling a complete story for the purpose of improving something [54]. Metrics are frequently indirect measurements and are very common in the measurement of SDOs. The following are some examples of metrics that can be collected for an SDO.

- SLOC - The number of Source Lines of Code
- NOM - The Number of Methods per class

- Complexity - A numerical measure of the code complexity (some common examples are McCabe [60] and Halstead [33])
- Design - The amount of coupling and cohesion present in the software code
- Source Code Analysis - Tools that determine whether code adheres to specified set of rules. Common examples are PMD[3] and FindBugs^{TM} [5], [17].

All these metrics are beneficial, but none of them tell the story of the entire SDO. Most of the metrics for an SDO, as seen in the list above, focus on the source code and development phase. Since metrics are indirect, it can be very difficult to match SDO performance with a metric or series of metrics. Metrics are great for tracking, but decision making based upon metrics alone is difficult. That is why many of the other techniques build upon metrics to provide a more complete overall picture of an SDO. Metrics are a great starting point, but more is needed to properly evaluate performance.

3.2 INDICATORS

Another common measurement technique is indicators. An *indicator* is simply a performance measure. Typically, a number of indicators will be placed together and displayed in some report or on some dashboard. Indicators can be crucial measurements within any business setting, and an SDO is no exception. Determining the correct indicators for an organization can be difficult, and many organizations incorrectly classify the indicators [67]. The differences between the indicators will be explored and possible measures for each indicator in an SDO will be presented. The four categories of indicators important to an SDO are shown in Table 1.

Performance Measure	**Description**
RI (Result Indicator)	What has been done?

[3]PMD is a source code analysis product. It is not an acronym.

Performance Measure	Description
KRI (Key Result Indicator)	How you have done?
PI (Performance Indicator)	What to do?
KPI (Key Performance Indicator)	How to dramatically increase performance?

Table 1: INDICATORS

Indicators can be used in just about every organizational setting from businesses to non-profit organizations. They are not unique to SDOs, and the exact indicators to track are very specific to the organization. The indicators chosen by one organization might not be the same as the indicators chosen by another organization. The following sections will explain the type of indicators in more detail and provide some examples for an SDO.

3.2.1 RESULT INDICATORS (RI) FOR AN SDO

Result Indicators are performance measures that summarize activity. All financial performance measures are result indicators. Result indicators are measured on a timely basis (daily, weekly, monthly) and are the result of more than one activity. They do not tell staff what needs to be done to improve the RI. For an SDO, some possible RIs are seen in Table 2.

Result Indicators
Requirements Implemented Per Month
Monthly SLOC
New Weekly Users
Monthly Development Hours
Webpage Views Yesterday
Monthly Development Hours
Monthly Server Uptime

Result Indicators
Quartely Software Sales

Table 2: RESULT INDICATORS FOR AN SDO

3.2.2 KEY RESULT INDICATOR (KRI) FOR AN SDO

Key Result Indicators are measures of multiple activities that give a clear picture of whether the organization is traveling in the right direction. Unfortunately, KRIs are commonly mistaken for KPIs [67]. KRIs do not tell an organization what is needed to improve the results. KRIs are highly beneficial for high-level management and not necessarily beneficial for staff working directly on the software. For an SDO, some possible KRIs are seen in Table 3.

Key Result Indicators
Customer Satisfaction
Net Profit
Money Spent on Fixing Software
% of New Features vs. Fixes
Time on Website
% Servers Meeting the Expected Availability

Table 3: KEY RESULT INDICATORS FOR AN SDO

3.2.3 PERFORMANCE INDICATOR (PI) FOR AN SDO

Performance Indicators are non-financial performance measures that help a team align themselves with the organizations strategy. These are important to the organizations success, but they are not the key measures that will lead to drastic improvement. They are

specifically tied to a team and all staff understand what actions need to be taken to improve the PI. For an SDO, some possible PIs are seen in Table 4.

Performance Indicators
% Test Coverage[4]
# of Project Defects from Key Customers
Requirements Scheduled for Next Release
# of Missed Requirements
% of Late Projects

Table 4: PERFORMANCE INDICATORS FOR AN SDO

3.2.4 KEY PERFORMANCE INDICATOR (KPI) FOR AN SDO

Key Performance Indicators are performance measures focusing on critical aspects for current and future organizational success. Notice, KPIs are not focused on historical performance, and they clearly indicate how to drastically increase performance. KPIs allow a team to monitor current performance and quickly take action to correct future performance. KPIs cover a shorter time frame than KRIs. KPIs consist of the following seven characteristics [67].

1. Not financial
2. Measured frequently (hourly, daily, weekly)
3. Acted on by CEO[5] and/or upper-level management
4. Clearly indicate the action required
5. Tie responsibility to a particular team

[4]Test Coverage is simply the percentage of the code that is being tested. Ideally, this number would be 100%, but higher is better.

[5]Chief Executive Officer

6. Have a significant impact

7. Encourage appropriate action

Antolic in [2] made one of the earliest attempts to identify and measure the KPIs for an SDO. Antolic's strategy focused around six KPIs.

1. Schedule adherence

2. Assigned content adherence

3. Cost adherence

4. Fault slip through

5. Trouble report closure rate

6. Cost per defect

However, according to the definition of KPI just presented, they are not really KPIs but instead KRIs. That is because none of the six clearly indicate the action required to improve the measure. Also, it is unclear if improving any of the measures will drastically improve performance.

For an SDO, Table 5 identifies some more appropriate KPIs.

Key Performance Indicators
Projects more than 20% behind schedule
Servers currently unavailable
Automated tests failing for more than 24 hours
Projects with test coverage less than 60%
Projects with more than 10 SIT defects
Unfixed, high priority PROD defects older than 1 week

Table 5: KEY PERFORMANCE INDICATORS FOR AN SDO

3.3 BALANCED SCORECARD

Developed in 1992 by Robert S. Kaplan and David P. Norton, the *balanced scorecard* is a set of measures that give management a quick and comprehensive view of the organization [52]. Originally created as an extension to the already existing financial measures, the balanced scorecard expanded the measures to include: customer focus, internal process, and learning/growth. This gave organizations a more comprehensive view that was not strictly financial. It allows an organization to focus on long-term strategic goals instead of just short-term goals. As a result of the strategic focus, the balance scorecard rapidly gained widespread adoption among businesses [53].

In 2010, David Parmenter [67] added two more characteristics to the balanced scorecard: employee satisfaction and environment/community. This results in a total of six characteristics for the balanced scorecard.

1. Financial
2. Customer Focus
3. Internal Process
4. Learning and Growth
5. Employee Satisfaction[6]
6. Environment/Community[6]

Balanced scorecards are great for easily displaying the important information about an organization. The downside is a balanced scorecard does not specify what exactly needs to be tracked. It can be very difficult to determine exactly what PIs, RIs, KRIs and/or KPIs to track in a balanced scorecard. It specifies six broad categories. It is also not specific to an SDO and it does not produce a single number. However, any new

[6] Added later by Parmenter [67].

measurement technique for an organization should be compared with the balanced scorecard.

3.4 PROJECT MANAGEMENT MEASUREMENT

A final strategy to measure an SDO is focused on the aspect of project management. Project management is the guidance applied to a project to ensure an effective and efficient completion. Proper project management will ensure all steps of the SDLC continue to progress and all obstacles are handled in a timely fashion. According to Putnam and Myers in [68], the five core measurements for managing software projects are:

1. **Quantity of function** - usually measured in terms of size (such as source lines of code), that ultimately execute on the computer
2. **Productivity** - as expressed in terms of the functionality produced for the time and effort expended
3. **Time** - the duration of the project in calendar months
4. **Effort** - the amount of work expended in person-months
5. **Reliability** - as expressed in terms of defect rate (or its reciprocal, mean time to defect)

A *process productivity* number is calculated based entirely on aspects of the SDLC and the five core measurements. It is a number targeted at project teams working on the SDLC.

3.5 A SIMPLER MEASUREMENT

It is important to note that SDOs do not just develop software. An SDO has many other duties including: deploying software, installing server hardware/software, writing

documentation, surveying users, research, innovation, education and other common business duties. Thus it is important to measure as many duties as possible.

How can the PIs, RIs, KRIs, and KPIs be combined to form a single value called the CRI (Cumulative Result Indicator)? If the indicators are targeted for upper management to understand performance, then KPIs are not the correct indicators. KPIs are targeted towards immediate action and future performance. RIs and KRIs are the most beneficial for upper management to gauge how an organization is doing. However, with so many possible RIs and even KRIs, it can be tricky to gain a quick understanding. The next section will present and explain a technique to combine KRIs into a single number for immediate and effortless evaluation of an SDO.

4 CUMULATIVE RESULT INDICATOR (CRI)

SDOs struggle to measure overall performance. The *Cumulative Result Indicator (CRI)* is an algorithm to provide a single number score to measure the performance of an SDO. It works by statistically analyzing the past performance of the organization and using that information to score an organization on current performance. CRI is a collection of the following five elements, which are actually KRIs, for an SDO.

1. **Quality**
2. **Availability**
3. **Satisfaction**
4. **Schedule**
5. **Requirements**

A separate CRI score is calculated for each element and then aggregated together to form an overall CRI score. A new CRI score will be calculated based upon the selection of a given time period (weekly, monthly, quarterly). CRI is not meant to be comparative

between organizations, but to measure the amount of increase or decrease a single organization exhibits across elements. CRI is made to be easily expandable to other elements if desired.

The scores for CRI will range from -1, indicating the worst performance, all the way to +1, indicating perfection. A score of 0 is an indication of meeting the basic expectations. A negative score indicates worse-than-expected performance and a positive score indicates better-than-expected performance. For example, a CRI score of 0.35 means the organization is performing 35% better than expected. Conversely, a score of -0.15 means an organization is performing 15% worse than expected.

Below are the attributes of the CRI scoring.

- The range of scores must have equal values above and below 0.
- The minimum score must equate to the worst possible performance, however that is defined.
- Similarly, the maximum score must equate to the best possible performance.
- A score of 0 must be average (or expected) performance.
- All individual elements must have the same scoring range.

As long as those 5 features are met, the range of scores can be anything. The range of $[-1, 1]$ was chosen because it is easy to scale to a different range such as $[-10, 10]$ or $[-100, 100]$. Thus scaling can be applied to obtain values in any appropriate range. The scaling factor is denoted with the variable k. The scale must be the same for all five elements.

4.1 ELEMENTS OF CRI

Each of the five elements of CRI has its own set of data that needs to be collected and formula for calculating a score. These five elements will be outlined in the next sections.

4.1.1 QUALITY

Measuring quality is a crucial part of accessing software development results. Poor quality means time, money, and resources are spent fixing the problems. As a result, new features are not being created. One of the key indicators of software quality is defects. It is important to measure the number of defects associated with a software release because industry-wide the current defect removal rate is only about 85% and this value should be increased to about 95% for high quality software [45]. Organizations are leaving too many defects unfixed. If organizations could lower the number of defects, then not as many defects would need to be fixed, which in turn would raise the defect removal rate.

Another aspect of defects is severity levels. A severity level indicates the importance of a defect that has been discovered. Although an organization can choose whatever severity levels they choose, it is common practice to use five severity levels [70]. The most severe level for a defect is 1. All other levels drop in severity from that point. Table 6 describes the five levels for defect severity.

Level	**Description**
1	Software is unavailable with no workaround
2	Software performance degraded with no workaround
3	Software performance degraded but workaround exists
4	Software functions but a loss of non-critical functionality
5	Others: minor cosmetic issue, missing documentation

Table 6: SOFTWARE DEFECT SEVERITY LEVELS

QUALITY DATA In order to properly score the quality of an SDO, certain data needs to be obtained in order to measure performance. Table 7 identifies the columns of data that will be used to create a score for the quality element of CRI. Each column is classified as

required or *optional*. This is to allow some flexibility in the model for organizations that collect varying amounts of data.

Column Name	Data Type	
Application ID	String (factor)	Required
Frequency Date	Date	Required
Development Effort	Integer	Required
Testing Effort	Integer	Optional
SIT Defects	Integer	Optional
UAT Defects	Integer	Optional
PROD Defects	Integer	Required

Table 7: QUALITY DATA NEEDED FOR CRI

The development and testing effort can come from any of the following choices for effort. It is possible that other measures will work for effort.

Actual Time This number is a representation of the total amount of time spent on a project. This number can be measured in any unit of time: hours, days, weeks, etc. Actual time can be applied to development or testing effort.

Estimated Time This number is a representation of the initial estimated amount of time spent on a project. This number can be measured in any unit of time: hours, days, weeks, etc. Estimated time can be applied to development or testing effort. It is common for the estimated and actual times to be different.

SLOC This number is the count of the total number of lines of source code for a project. Obviously, this item only counts as a level of effort for development unless coding in used to generate automated testcases. [7]

[7] Automated testing is the process of creating software to automatically run tests against other software. The adoption of automated testing is varied and it is not a solution in all cases [69].

Modified Lines Of Code This number is a count of the number of modified lines of source code. Modified lines is defined as the number of deleted, added, and modified lines of source code. This number is different from above since it does not include all the lines of source code. Similar to above, this number makes more sense for development effort.

Testcases A testcase is a step or series of steps followed to validate some expected outcome of software. Organizations will create a number of testcases to be validated for a software system. The number of such testcases could be used as a level of testing effort.

Notice the data does not include a severity level. The severity level should be handled before being stored. A good technique is to count the defects based upon the weighting scheme in Table 8 [70]. For example, finding one defect of severity level 5 will result in a total count of one. However, finding one defect of severity level 2 will result in a total count of 15. This strategy helps to standardize the number of defects found. An organization can alter the values of Table 8 based upon priorities or use a different technique if desired. It is important to establish a standard, meaningful number for SIT defects, UAT defects, and PROD defects which manages severity appropriately.

Severity Level	Weight
1	30
2	15
3	5
4	2
5	1

Table 8: DEFECT SEVERITY LEVEL WEIGHTING

QUALITY FORMULA The first step in creating a score for the quality element is analysis of the historical data to create a baseline function. The historical data is all quality data collected before a given point in time. Some common historical cutoffs are the current date or the end of the previous fiscal year. Then a mathematical model, called the *baseline quality function*, to predict PROD Defects will be produced. In statistical terms, the response is *PROD Defects* and the predictors are: *UAT Defects*, *SIT Defects*, *Testing Effort*, and *Development Effort*. Some of the following strategies to find a reasonable model include:

- Removal of outliers and/or influential points
- Linear Regression
- Stepwise Regression
- Ridge Regression for suspected multicolinearity

Once a model has been found, it will be labeled as f and it will not change. The function f can be the same for all Application IDs or it can be different for each Application ID or any combination of Application IDs. It serves as the quality baseline for CRI. All future quality scores will be dependent upon the original f. Once set, the model will not change.

After the model f has been determined, it is time to calculate the quality score for each application ID within the given time period. The quality score for each Application ID can be calculated as follows.

$$S_{1_i} = \begin{cases} \text{where } f_i \geq d_i & : \frac{f_i - d_i}{f_i} \cdot k \\ \text{where } d_i > f_i & : \frac{f_i - d_i}{\sigma_i^2} \cdot k \end{cases} \text{, calculate quality score for each app } i$$

where

- S_{1_i} is the quality score for Application ID i
- k is the scaling factor to produce results in the range $[-k, k]$
- n is the number of Application IDs
- d_i is the actual PROD defects for Application ID i
- f_i is the function to predict PROD defects for Application i based upon *UAT Defects*, *SIT Defects*, *Testing Effort*, and *Development Effort*
- σ_i^2 is the estimated variance for Application i

Then the overall quality score is calculated as below.

$$S_1 = \sum_{i=1}^{n} w_i S_{1_i}$$

where

- S_1 is the combined quality score for all Application IDs, a weighted average
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$

Then S_1 represents the CRI quality score for that given time frequency.

4.1.2 AVAILABILITY

All the new requirements and great quality do not matter if the software is not available. Thus it is essential to set an expected Service Level Agreement (SLA)[8] and measure performance against that SLA. The following section will outline the data needed to properly calculate an SLA and to calculate the CRI score for availability.

[8]For an SDO, the SLA is a contract specifying the amount of time software will be available during a given time period.

Special Note: The Service ID for availability does not have to be the same as the Application ID for quality or any of the other elements. Some organizations have a one-to-one mapping between Applications being developed and services being deployed. Others have more complex scenarios that require multiple applications to be combined to form a service. Then the availability of the system is tracked.

AVAILABILITY DATA Table 9 identifies the necessary data to calculate the CRI element score for availability. Notice the three optional fields: *Uptime*, *Scheduled Downtime*, and *Unscheduled Downtime*; they are optional because they can be used to calculate the *Percent Uptime*. The *Percent Uptime* is the important value for the CRI schedule score. Here are the two common approaches for calculating percent uptime:

The preferred method:

$$\text{Percent Uptime} = \frac{\text{Uptime}}{\text{Uptime + Scheduled Downtime + Unscheduled Downtime}}$$

and the alternative method:

$$\text{Percent Uptime} = \frac{\text{Uptime}}{\text{Uptime + Unscheduled Downtime}}$$

The only difference is the removal of scheduled downtime from the calculation. The calculation approach is typically specified in the contract associated with the SLA. Thus, the Percent Uptime is important and it can either be supplied in the data or calculated from the optional fields. For the purposes of this dissertation, the percent uptime should be handled in decimal form and not as a percent out of 100.

Column Name	**Data Type**	
Service ID	String	Required
Frequency Date	Date	Required
Uptime	Float	Optional

Column Name	Data Type	
Scheduled Downtime	Float	Optional
Unscheduled Downtime	Float	Optional
Percent Uptime	Float	Required
Expected Percent Uptime	Float	Required

Table 9: AVAILABILITY DATA NEEDED FOR CRI

AVAILABILITY FORMULA The formula for availability is more straightforward than the quality formula. It does not include any analysis of the historic data. That lack of historical analysis is avoided since the SLA provides an existing baseline to measure against. The following formula is simply a percentage the SLA was exceeded or missed.

$$S_{2_i} = \begin{cases} \text{where } A_{a_i} \leq A_{e_i} & : \frac{A_{a_i} - A_{e_i}}{A_{e_i}} \cdot k \\ \text{where } A_{a_i} > A_{e_i} & : \frac{A_{a_i} - A_{e_i}}{1 - A_{e_i}} \cdot k \end{cases}, \text{ calculate availability score for each sys } i$$

where

- S_{2_i} is the availability score for System ID i
- k is the scaling factor to produce results in the range $[-k, k]$
- A_{a_i} is the actual availability for System ID i
- A_{e_i} is the expected availability for System ID i

Then the overall availability score is calculated as below.

$$S_2 = \sum_{i=1}^{n} w_i S_{2_i}$$

where

- S_2 is the combined availability score for all Service IDs, a weighted average
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$

Then S_2 represents the CRI availability score for that given time frequency.

4.1.3 SATISFACTION

The satisfaction of users, customers, and/or business partners is the third element to be measured. This element is important because in an established business, retaining customers is less expensive than attracting new customers [4]. Depending upon the type of SDO, the customers may be internal or external to the organization. For the remainder of this section, the term customer will be used to represent any person who is responsible for guidance, decision-making or use of the software. The term customer can refer to a: user, paying or nonpaying customer, internal or external business partner, or any other person deemed influential to the development of the software.

If one element of CRI was to be rated as the most important, satisfaction would be it. Without satisfied customers, the rest of the measures do not matter. For example, having a quality application that is always available does not matter if the application is not what the customer wants.

Surveys are used to measure satisfaction for CRI. A series of statements will be presented to all or a subset of the customers. Any customer that chooses to respond to the survey is considered a respondent. A respondent can rate statements based upon a Likert Scale[9] with a numerical response where the minimum value indicates maximum disagreement and the maximum value indicates the maximum agreement. Common rating scales would be from 1 to 5 or from 1 to 3. An example survey can be seen in Table 10.

[9]"The Likert Scale presents respondents with a series of (attitude) dimensions, which fall along a continuum." [18]

ID	Statement	Disagree	Neutral	Agree
1	I find the software easy to use.			
2	I would recommend this software to others.			
3	The software makes me more productive.			
4	I am happy with this software.			

Table 10: SAMPLE SURVEY FOR SATISFACTION

ISSUES WITH SURVEYS Surveys present a number of challenges that need to be presented and briefly discussed. Here are some of the issues that need to be addressed when using surveys.

Text

The specific text used in the questions or statements is very important. The text cannot be too vague. Also, the text must be clear enough to eliminate misinterpretation. Survey questions must be complete and not include gaps. For example, if an age range is presented, it must include all possible ages. These are just some of the difficulties with getting the text correct in surveys.

Number and Ordering

The number of questions is important. Too many questions and the respondents will lose interest and begin responding without the adequate attention needed. Plus, if the survey is too long there is the risk of quitting before completion. A short survey might not cover the adequate amount of material. Both short and long surveys run the risk of providing inaccurate responses. After determining the best number of questions, the ordering of the questions is important. Previous survey questions can have an unintended impact on responses. Thus, the ordering of questions needs to be addressed.

Sampling

Next is the issue of sampling. Not every customer can be surveyed, so sample sets of customers need to be presented with a survey. In the case of CRI, there are two possible scenarios for sampling.

1. When an SDO is part of a larger organization, there typically is a small number of business partners that help to guide and direct the work performed by the SDO. In this case, the business partners might be the the ones offering survey responses and they should all be willing to respond. Thus, those business partners represent the entire population, and the surveys should result in a 100% response rate which is technically a census. The only bias that will be present here is the bias of the business partners and sampling cannot control for that.
2. End-users will be surveyed for satisfaction. Obviously, the entire population cannot be surveyed, so a probability sample should be randomly created. Even then, bias will be present.
    - Not all users will respond. This is because survey respondents tend to sit at the extremes of either satisfied or dissatisfied. Thus the results will tend to indicate that separation.
    - Even with probability sampling it is possible to miss entire groups of population members. For example consider a banking application such as a savings account, a survey would most likely be presented online, and it would have a coverage bias due to the exclusion of savings account holders that do not bank online.
    - A selection bias can occur when some members of the population have a higher probability of inclusion in the sampling frame than others. One example could be a user with multiple savings accounts. The selection

bias is typically easy to avoid if the bias is identified. Weighting is a common solution for selection bias.

For more on creating appropriate surveys, see [77] by Snijkers, Haraldsen, Jones, and Willimack. They present a framework named generic statistical business process model (GSBPM) for conducting surveys in a business or organizational setting. GSBPM covers the issues above as well as a few more issues such as response storage and risks. Also, Cowles and Nelson provide another good resource for preparing and conducting surveys [18]. They even include entire chapters on both writing survey questions and survey errors.

SATISFACTION DATA Once the surveys have been distributed and the results collected, Table 11 displays the data that needs to be collected in order to calculate the satisfaction element score for CRI.

Column Name	**Data Type**	
Question ID	String	Required
Question Text	String	Optional
Respondent ID	String	Optional
Frequency Date	Date	Required
Response	Integer	Required
Response Date	Date	Optional
Application ID	String	Optional

Table 11: SATISFACTION DATA NEEDED FOR CRI

SATISFACTION FORMULA After collecting the necessary survey data from Table 11, calculating the score is rather straightforward. The scores for each question are averaged and then those values are averaged together. If some survey questions are more important

than others, the formula could be easily modified to include weighting.

First the score for each question needs to be calculated.

$$S_{3_i} = k \cdot \frac{\sum_{j=1}^{m} (\frac{2a_{ij} - min - max}{max - min})}{m}$$

- S_{3_i} is the satisfaction score for Question ID i
- k is the scaling factor to produce results in the range $[-k, k]$
- a_{ij} is the answer to question i for respondent j
- n is the number of questions
- m is the number of respondents
- min is the minimum score for a question
- max is the maximum score for a question

Then the satisfaction score is calculated as below. Use a weighted average to combine the question scores:

$$S_3 = \sum_{i=1}^{n} w_i S_{3_i}$$

where

- S_3 is the combined satisfaction score for all Questions IDs, a weighted average
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$.

Then S_3 represents the CRI satisfaction score for that given time frequency.

4.1.4 SCHEDULE

Delivery of software in a timely manner is an essential part of being a successful SDO. Being able to meet scheduled deadlines is a sign of accurate estimation and planning. Drastically missing deadlines is a sign of an SDO with a process that needs refinement. Studies have shown that software projects exceed the estimates by an average of 30% [50]. Thus it is important to score SDOs on accurate schedule adherence. Without tracking and measuring schedule adherence, it will not improve.

The CRI schedule score provides a numeric value to indicate the amount schedules are missed or exceeded. The score provides a cumulative measure of the performance as compared to other months. The score is based upon the historical deviance of estimates for projects. Projects completing on time will be given a score of 0. Projects finishing early will be rewarded with positive scores increasing toward k. Alternatively, late projects will be given negative scores that approach $-k$ as the projects become more late.

SCHEDULE DATA In order to calculate the schedule score, certain dates need to be present. Table 12 outlines the necessary data for schedules. One date is considered optional as it is not used in the CRI calculation, but it is an important date that could be useful for future enhancements to CRI.

Column Name	**Data Type**	
Project ID	String	Required
Frequency Date	Date	Required
Scheduled Start Date	Date	Required
Scheduled Finish Date	Date	Required
Actual Start Date	Date	Optional
Actual Finish Date	Date	Required

Table 12: SCHEDULE DATA NEEDED FOR CRI

SCHEDULE FORMULA Schedule has a clear date for finishing on-time, however there are not clear bounds as to how early or late a project can be delivered. Thus, the formula for schedule is more involved than availability or satisfaction. It requires some analysis of the historical data. The first step of the formula is determining how often projects are early or late, and by how much a project is early or late. This can be accomplished by looking at the distribution of the data. Specifically, look at what percentage of the entire project duration the schedule was missed.

$$\Delta_i = \frac{F_{a_i} - F_{s_i}}{F_{s_i} - B_{s_i} + 1}$$

where

- F_{a_i} is the actual finish date of project i
- F_{s_i} is the scheduled finish date of project i
- B_{s_i} is the scheduled beginning date of project i
- Δ_i is the proportion the schedule was missed for project i

Once all the Δ_i's have been determined, a distribution must be fit to the data. There are several techniques for testing the fit of a distribution: histograms, chi-square, Kolmogorov-Smirnov, Shapiro-Wilk, or Anderson-Darling [20], [55]. The distribution is needed for the Cumulative Distribution Function (CDF) . The CDF maps the values to a percentile rank within the distribution [23]. The CDF will be transformed to create the schedule score for CRI. Since all CDF functions fall within the range $[0, 1]$, the function needs to be shifted to center around 0, and then doubled to fill the desired range of $[-1, 1]$. Thus the CRI schedule score for each project becomes the following.

$$S_{4_i} = 2k \cdot \left(CDF(\Delta_i) - \frac{1}{2} \right)$$

where

- S_{4_i} is the schedule score for Project ID i
- k is the scaling factor to produce results in the range $[-k, k]$

Then the overall schedule score is calculated as below.

$$S_4 = \sum_{i=1}^{n} w_i S_{4_i}$$

where

- S_4 is the combined schedule score for all Project IDs, a weighted average
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$

Then S_4 represents the CRI schedule score for that given time frequency.

ALTERNATE APPROACH An alternative approach for scoring schedule goes as follows. The best possible score should be achieved when meeting the estimated date exactly. The maximum score should come from the best estimate. Then given historical release data, it is easy to determine an average Δ between the actual and the estimated. Finishing a project within that Δ should result in a positive score. Outside the Δ results in negative scores. For example, a project releasing one day early or one day late would receive the same score because in both cases the estimate was missed by one day.

The first step of the formula is finding the percentage the schedules were missed for historical projects. The calculation treats over- and under-estimating the schedule the same. The same penalty is applied in both cases. For example, being 15% late will result in the same score as being 15% early. Perform this calculation only for projects that did not exactly meet the estimated finish date.

$$\Delta_i = \left| \frac{F_{a_i} - F_{s_i}}{F_{s_i} - B_{s_i} + 1} \right|$$

Find the average of the Δ_i's. This is the average proportion of a missed schedule.

$$\bar{\Delta} = \frac{\sum_{i=1}^{n} \Delta_i}{n}$$

The formula for schedule is then a percentage above or below the Δ. The number is calculated for each project, and then averaged to form the schedule score.

After $\bar{\Delta}$ is calculated, the following formulas are used to create the schedule scores for each project and then the averaged schedule score.

$$S_{4_i} = \begin{cases} \text{where } \Delta_i \geq 1 & : -1 \cdot k \\ \text{where } \Delta_i \leq \bar{\Delta} & : \frac{\bar{\Delta} - \Delta_i}{\bar{\Delta}} \cdot k \\ \text{where } \Delta_i > \bar{\Delta} & : \frac{\bar{\Delta} - \Delta_i}{1 - \bar{\Delta}} \cdot k \end{cases} \text{, calculate schedule score for each project } i$$

$$S_4 = \sum_{i=1}^{n} w_i S_{4_i}$$

where

- S_4 is the combined schedule score for all Project IDs, a weighted average
- k is the scaling factor to produce results in the range $[-k, k]$
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$
- n is the number of projects
- F_{a_i} is the actual finish date of project i
- F_{s_i} is the scheduled finish date of project i
- B_{s_i} is the scheduled beginning date of project i

- Δ_i is the percent the schedule was missed
- $\bar{\Delta}$ is the average percent schedules are missed
- S_{4_i} is the schedule score for project i

Then S_4 represents the CRI schedule score for that given time frequency. Again, this was just an alternate approach to scoring schedule. It will not be used in the case studies.

4.1.5 REQUIREMENTS

The requirements of an SDO are important. Requirements are desired new features or enhancements to a software product. It is important to know how many requirements were scheduled to be completed versus how many actually got completed. They provide a measurement of the amount of work being completed. However, not all requirements are created equal. Some requirements can be quickly and easily implemented while other requirements will take much longer. It is often difficult to know the challenges ahead of time. Due to this uncertainty, estimating the number of requirements a team can complete in a given time frame can be difficult. Thus, the number of requirements is one way to measure amount of work but there are others. Here are three possible choices in descending order of preference.

1. **Function Points** - Function points measure the size of the software functionality, not the size or time required to implement the functionality [43]. If done properly, function points will provide the most concise measurement for amount of work being completed.

2. **Story Points** - Story points, sometimes just referred to as stories, are broken down requirements that can be completed in an Agile Sprint of two to four weeks. A story is usually smaller, simpler, and more concise that a plain requirement. Unfortunately, these only apply to an SDO using the Agile methodology, specifically Scrum [3].

3. **Requirements** - These are just the requirements as written by the users or business partners. Each requirement can vary greatly from another requirement. Estimating number of requirements can be a difficult if not impossible task.

To summarize, this element is not completely a measurement of requirements but rather a measurement of the work being completed.

REQUIREMENTS DATA Table 13 specifies the data required to compute a score for the requirements element of CRI. Again, scheduled requirements and actual requirements can also be a count of story points or function points. All of the columns are required. The data collected is the frequency date and then the number of requirements scheduled to be completed and the actual number of requirements completed.

Column Name	**Data Type**	
Project ID	String	Required
Frequency Date	Date	Required
Scheduled Requirements	Integer	Required
Actual Requirements	Integer	Required

Table 13: REQUIREMENTS DATA NEEDED FOR CRI

REQUIREMENTS FORMULA The requirements formula is the percentage above or below the scheduled number of requirements. Requirements have a nice lower bound of 0 since negative requirements cannot be completed[10]. Unfortunately, an upper bound does not exist. The variability of requirements is not as large as the variability of the number of defects, so a simpler strategy can be used. For requirements, a multiplier b will be used to find the upper bound. The number of scheduled requirements should be multiplied by b to obtain the upper bound. The value of b should be determined by looking at the historical

[10]Theoretically, a negative requirement would be considered a defect.

data to make sure no number of completed requirements will exceed b times the number of scheduled requirements. An example will be shown in Section 6.4. Common choices for b will be 1 and 2. Also, the formula will be created to deal with a value going above the upper bound.

$$S_{5_i} = \begin{cases} \text{where } R_{a_i} > R_{s_i} \cdot (b+1) & : 1 \cdot k \\ \text{where } R_{a_i} \leq R_{s_i} & : \frac{R_{a_i} - R_{s_i}}{R_{s_i}} \cdot k \\ \text{where } R_{a_i} > R_{s_i} & : \frac{R_{a_i} - R_{s_i}}{b \cdot R_{s_i}} \cdot k \end{cases}, \text{ requirements score for project } i$$

where

- S_{5_i} is the requirements score for Project ID i
- k is the scaling factor to produce results in the range $[-k, k]$
- R_{a_i} is the actual requirements completed for Project ID i
- R_{s_i} is the expected requirements completed for Project ID i
- b is the multiplier to determine the upper bound

Then the requirements score is calculated as below. Use a weighted average to combine the requirements scores from the Project IDs.

$$S_5 = \sum_{i=1}^{n} w_i S_{5_i}$$

where

- S_5 is the combined requirements score for all Project IDs, a weighted average
- $w_i > 0$ for all i
- $\sum_{i=1}^{n} w_i = 1$

Then S_5 represents the CRI requirements score for that given time frequency.

4.1.6 OVERALL CRI SCORE

In order to accomplish the single number score that CRI requires, the five element scores must be combined. The combination of the scores is a weighted average. The weights can be set based upon the priority of the SDO. Thus, the overall CRI score is calculated as below.

$$CRI = \sum_{i=1}^{n} w_i S_i \text{ where } \sum_{i=1}^{n} w_i = 1$$

where

- CRI is the overall CRI score for the time frequency

This weighted average allows for a score to be computed even when not all five elements are present. Just perform the weighted average with as many elements as present. This technique allows CRI to be implemented before data for all the elements has been collected. Just begin to average elements as they become accessible.

4.2 CORRELATIONS IN CRI

It is possible that two or more of the five elements of CRI will be correlated. This means that one of the elements can be predicted based upon the values of the other elements. Although it is possible for correlation to occur between any of the elements, the satisfaction element is an obvious element which deserves attention due to the human involvement of the surveys. If a schedule is missed or an important requirement dropped, that could have a large negative effect on the satisfaction surveys. The same could be said of quality or availability with regard to the satisfaction. However, satisfaction is not the only potentially correlated element. It is also possible that a decrease in quality could result in unexpected downtime which could have a negative result on availability. Similarly, if requirements are added, it is possible the schedule will be negatively

impacted. Also, if requirements are dropped, the quality might suffer due to missing functionality.

It is impossible to know which or if correlations will always exist. Thus it is necessary to check for correlations after determining element and overall CRI scores. If an element is determined to be correlated with another element, neither element should be dropped, but rather one of the elements should be weighted less than the other correlated element. This technique keeps the most data available but lessens the importance of the correlated element.

4.3 SENSITIVITY OF CRI

The sensitivity of the formulas should be tested as the scores should not fluctuate drastically for similar values. There are two possible techniques for testing the sensitivity of the formulas.

1. Given the historical data that has been collected, alter the values by some small random amount. Then recalculate the CRI element score and compare with the original score. This technique can be repeated many times in order to verify small changes do not largely affect the score. Thus, a formula which is not overly sensitive.

2. Another technique is to use Monte Carlo methods to randomly generate input values for the element functions. This can be done by finding the distributions of the historical data, and randomly selecting from that distribution. If historical data is not available, then the normal distribution can be used. See [76] for more information on sensitivity analysis in statistical modeling.

.

4.4 CRI COMPARED

CRI is one way to evaluate an SDO, and it is also a technique of software analytics. Therefore, it is beneficial to compare CRI with some of the other techniques and guidelines available. The next sections will provide those comparisons.

4.4.1 CRI VS. FOCUS AREAS OF SOFTWARE ANALYTICS

Earlier, in the introduction section 1.4.3, 3 main focus areas for software analytics were presented. Table 14 provides a explanation of how CRI addresses each focus area. As can be seen, CRI clearly addresses the three main focus areas of software analytics. CRI does not provide any mechanisms for improving the focus areas, but it provides a consistent mechanism to measure the focus areas.

Focus Area	**Why CRI?**
User Experience	One of the five elements of CRI is satisfaction. While not all of the questions focus solely on the user experience, the entire purpose of the survey is to determine if the user is satisfied with the software product. Does it have the correct features? Are new features added in a timely manner? Of course, specific survey questions can be created to focus solely on a certain user experience.
Quality	Again, one of the five elements specifically focuses on quality. CRI provides a single number to measure quality. Therefore, it is easy to track changes in quality over time. CRI does not address how to improve the quality, but without a consistent measurement, it would be impossible to determine the change in quality.

Focus Area	Why CRI?
Development Productivity	The combination of CRI elements, schedule and requirements, provide an indication of development productivity. The schedule element measures the productivity related to estimated schedule. Similarly, the requirement element measures the amount of work actually being completed.

Table 14: SOFTWARE ANALYTICS FOCUS AREAS AND CRI

4.4.2 CRI VS. IMPORTANT QUESTIONS OF SOFTWARE ANALYTICS

Also, section 1.4.3 mentions three important questions that software analytics must address. Table 15 presents the three questions and a description of how CRI addresses that specific question. It is clear that CRI addresses the questions. CRI is a beneficial technique of software analytics when applied to SDOs.

Question	Why CRI?
How much better is my model performing than a naive strategy, such as guessing?	CRI provides consistency which may not exist without it. Therefore, CRI removes the guesswork of measuring an SDO.
How practically significant are the results?	The CRI score is consistent and easy to comprehend. Thus comparison with past performance is quick and simple. This is a significant advantage for software development organizations.

Question	Why CRI?
How sensitive are the results to small changes in one or more of the inputs?	The question was extensively addressed in section 4.3. CRI is not overly sensitive to small changes in the inputs.

Table 15: IMPORTANT QUESTIONS FOR SOFTWARE ANALYTICS AND CRI

4.4.3 CRI VS. BALANCED SCORECARD

Section 3.3 discusses the characteristics of the balanced scorecard. Table 16 presents a comparison of the characteristics of a balanced scorecard versus CRI. As the newer two characteristics of a balanced scorecard have only existed since 2010 and the adoption is limited, the comparison will only be against the original four balanced scorecard characteristics.

Balanced Scorecard	CRI?	Explanation
Financial	No	CRI does not address financial as it is best suited for an organization that treats software development as a fixed, budgeted expense. If the budget is fixed, CRI provides a number to indicate the amount of value for that fixed budget.
Customer Focus	Yes	CRI includes a customer survey which is completely customer focused.
Internal Process	Yes	CRI is highly focused on internal processes. The CRI elements of schedule and requirement are completely focused on how reality meets the expected process. CRI is negatively impacted when internal processes are followed.

Balanced Scorecard	CRI?	Explanation
Learning/Growth	No	CRI does not address this characteristic.

Table 16: BALANCED SCORECARD VERSUS CRI

4.4.4 CRI VS. PROJECT MANAGEMENT MEASUREMENT

Previously in Section 3.4, the project management measurement was presented. Its greatest limitation is the lack of focus on the entire SDO. Project management measurement says nothing about the availability of the software infrastructure or the satisfaction of the users. It is not near as comprehensive as either the balanced scorecard or CRI. Actually, CRI incorporates the five core measurements from project management. Plus, the process productivity number is less suitable for upper management and more suitable for project teams. Although it does provide a single process productivity number, it is does not have the same focus as CRI.

5 SDLC ANALYTIC ENGINE

In order for an SDO to properly track the elements of CRI, a data storage system should be available to store the appropriate data. A consistent storage system should help to avoid the problem of inaccurate data caused by numerous manipulations of the existing data [66]. Plus, if the system is implemented correctly by allowing limited changes to existing data, it will be able to alleviate some of the dishonesty that is currently present in software projects [71]. This storage system will be named the SDLC Analytic Engine (SDLC-AE).

Once all the SDLC data is collected into a single place, there are many possible applications. Software analytics will be much easier to create and gamification will be much more easily attainable. CRI is just one possible application of the SDLC-AE. Figure 6 provides an overview of the data that could potentially be stored in the SDLC-AE as it

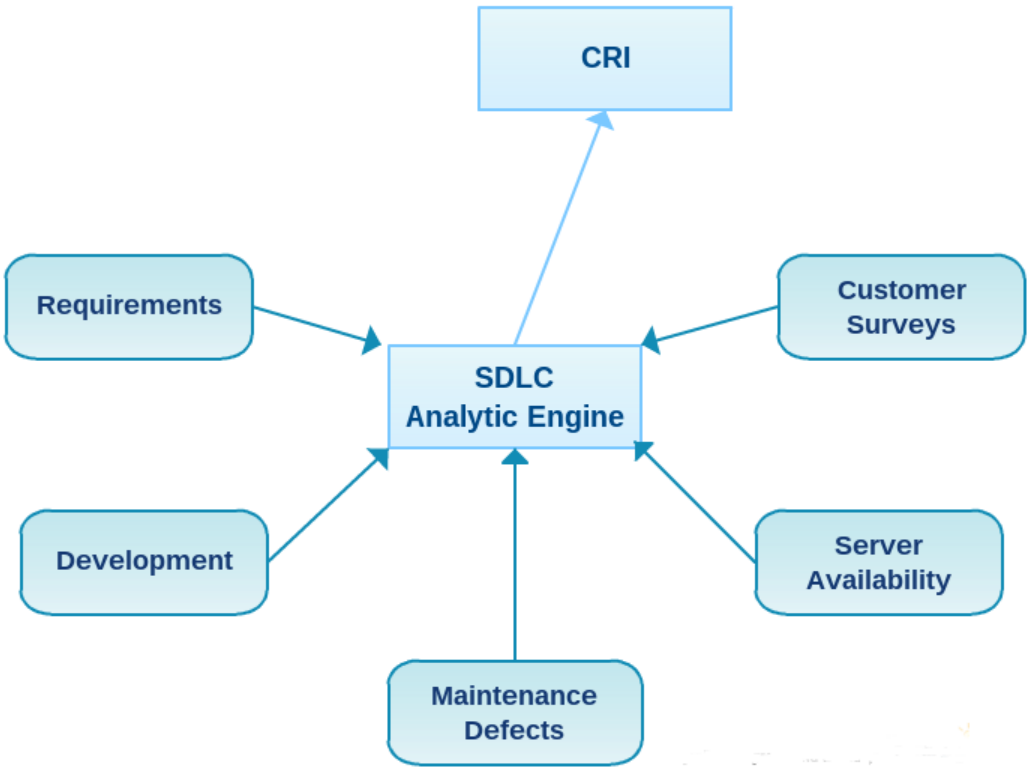

Figure 6: SDLC ANALYTIC ENGINE

relates to CRI. The SDLC-AE does not specify how the data is entered, just how the data is stored.

5.1 DATABASE STRUCTURE

All the data necessary to compute CRI needs to be stored in a database. This section will lay out the structure of tables and relationships necessary to store all the data within a relational database. The SQL[11] is written for an Oracle database, but the scripts can be modified to work with other databases such as: PostgreSQL, SQL Server, or MySQL.

5.1.1 TABLES FOR RAW CRI DATA

The first set of tables that need to be created are tables to store the raw data that is collected. These tables will match up with the necessary data for each of the five elements

[11]Structured Query Language (SQL) is a programming language designed for managing data in relational database management system.

of CRI. Figure 7 provides a visual description of the tables that are needed. The tables have no relationship with each other since they are raw data. The primary goal of these tables is to store the raw data in a single location. The five table names are:

1. QUALITY_RAW
2. AVAILABILITY_RAW
3. SATISFACTION_RAW
4. SCHEDULE_RAW
5. REQUIREMENTS_RAW

Notice these table names match with the five elements of CRI.

Appendix B.1 provides the necessary SQL statements to create the database tables for storing the raw CRI data.

5.1.2 INTERMEDIATE SCORE TABLES FOR CRI

The next set of tables that need to be created are for the intermediate level element scores. These tables hold the element scores at the application_id, service_id, and project_id level. Figure 8 provides a visual overview of the necessary tables. Furthermore, these tables also lack relationships between one another because at this point, all the element scores are still being treated independently. The five table names are:

1. QUALITY_SCORE
2. AVAILABILITY_SCORE
3. SATISFACTION_SCORE
4. SCHEDULE_SCORE
5. REQUIREMENTS_SCORE

Figure 7: TABLES FOR RAW CRI DATA

Again, these table names match very closely with the five elements of CRI.

Appendix B.2 provides the necessary SQL statements to create the database tables for storing the intermediate CRI scores.

5.1.3 FINAL SCORE TABLES FOR CRI

The final set of tables consists of only two tables.

1. ELEMENT
2. CRI_SCORE

Figure 8: TABLES FOR INTERMEDIATE CRI SCORES

The first table is the ELEMENT table. It simply stores the CRI element (Quality, Availability, Satisfaction, Schedule, Requirements, Overall) and an optional description. The second table is the CRI_SCORE table. It stores all the final element scores and the final overall CRI score. It is related to the ELEMENT table. Figure 9 provides a visual representation of the relationship between the two tables.

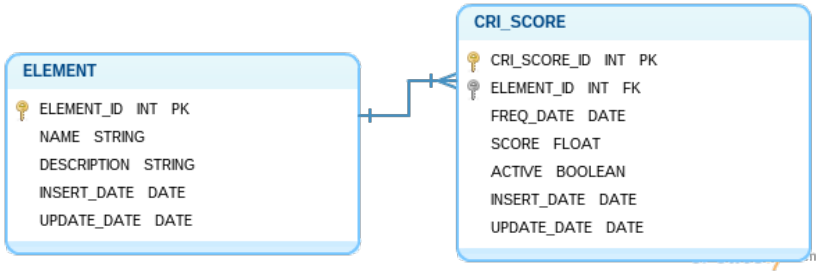

Figure 9: TABLES FOR FINAL CRI SCORES

Appendix B.3 provides the necessary SQL statements to create the database tables for storing the final scores for each element and the final overall CRI scores for each time frequency.

6 CASE STUDY: SCORING AN SDO OF A LARGE FINANCIAL INSTITUTION

Data has been collected from the software development processes of an SDO within a large financial institution[12]. The data collection was from 2007 to January 2015. Only 4 elements had available data, and not all elements had data available for the entire time period. The elements to be used are: quality, availability, schedule, and requirements. CRI is still effective when not all data is available. The overall CRI score will then be a weighted average of the available elements. This section will serve as a guide to preparing the CRI score based upon the available data.

After the data has been collected, the raw data can be stored in the SDLC if desired. Then scores for each of the 4 elements can be calculated. For this example, the value of k will be 100 and the time frequency will be monthly. Thus, scores will fall in the range $[-100, 100]$. Also, equal weighting is applied to all elements and all Project IDs, Application IDs, and System IDs.

6.1 QUALITY

The first step for dealing with the quality data is a quick analysis of the data. Table 17 provides some descriptive statistics for the quality data. Testing hours were not captured, but they are optional, so the analysis can continue.

Column	Min	Max	Median	Mean	Variance
Development Effort	0	26937	300	1637	18383464
Testing Effort	NA	NA	NA	NA	NA

[12] All the raw data files are available at [82]

Column	Min	Max	Median	Mean	Variance
SIT Defects	0	1106	1	45.86	24528
UAT Defects	0	277	0	10.28	1306
PROD Defects	0	1216	5	51.5	20311

985 obs. from 23 Application IDs from 2007 to 2015

Table 17: QUALITY DATA DESCRIPTIVE STATISTICS

Notice, the data for quality goes from October 2007 to January 2015 and contains 985 observations. This is important because historical data can be used to create the baseline quality function. All quality for the years 2007 through the end of 2013 will be used as historical data for the purposes of creating the baseline quality function. Once the historical data is separated, it results in 799 observations to be used for creating the baseline quality function. Figure 10 shows scatterplots of PROD_DFTS versus the independent variables of DEV_EFF, SIT_DFTS, and UAT_DFTS. It can be seen that some correlations exist between the variables.

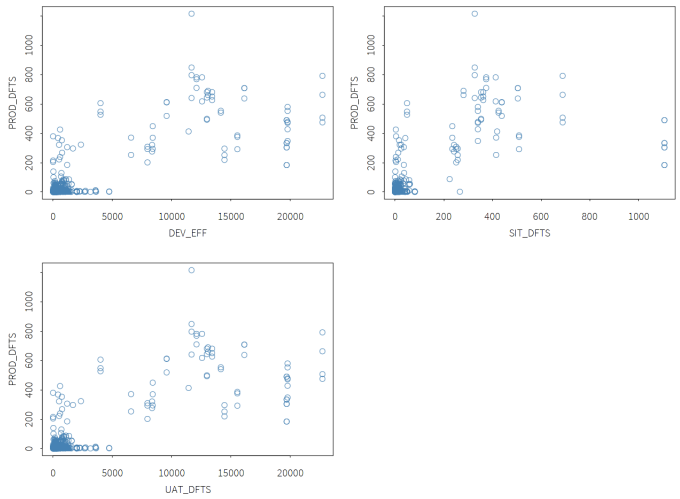

Figure 10: QUALITY DATA PLOTS: DEPENDENT VS. INDEPENDENT VARIABLES

Figure 10 also shows the presence of a possible outlier with 1216 PROD_DFTS.

That data point is dropped for the remaining analysis. At this point, a simple linear regression model can be fit to the remaining 798 observations. At this point, no transformations have been performed on the data. The linear regression model yields all 3 independent variables as significant and an overall $R^2 = .72$. The source code and some further analysis can be found in Appendix C.1. It appears to be a good fit and thus all Application IDs will have the same baseline quality function. Therefore, f does not need a subscript, and f can be seen below.

$$f = 5.92 + 0.035 \cdot DEV\_EFF - 0.36 \cdot SIT\_DFTS + 1.05 \cdot UAT\_DFTS$$

Now that f has been determined, the quality scores for each application can be calculated for all the months beyond 2013. Appendix C.3 provides the necessary R code to perform the calculations. Figure 11 shows the quality scores for 2014 and beyond. As can be seen, the scores are greatly above expectations. That is an indication of improved quality over historical performance.

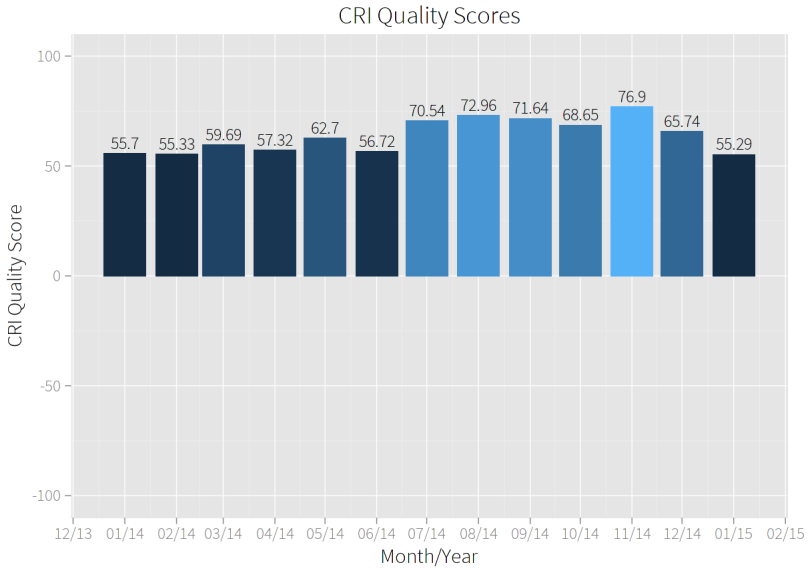

Figure 11: CRI QUALITY SCORES

6.2 AVAILABILITY

For the availability data, some descriptive statistics can be seen in Table 18. The percent uptime was previously calculated for this data so uptime, scheduled downtime, and unscheduled downtime are not needed. Availability does not rely upon analysis of historical data since a known upper and lower bound exist, 1 and 0 respectively.

The numbers in Table 18 indicate the values are very near 1. This is expected as SDOs set SLA uptime values near 1, and even strive to meet an uptime of 1. An uptime of 1 means the system was available the entire time period. Thus numbers for uptime near 1 are highly desirable for an SDO expecting to score well for availability.

Column	**Min**	**Max**	**Median**	**Mean**	**Variance**
Percent Uptime	0	1.0	1.0	0.9745	0.023
Expected Percent Uptime	0.93	1.0	0.98	0.9769	0.977

7522 obs. from 83 Application IDs from 2008 to 2015

Table 18: AVAILABILITY DATA DESCRIPTIVE STATISTICS

Since percent uptime has already been calculated and no historical analysis needs to be performed, the calculation of the scores can be performed. R code to compute the CRI availability scores can be found in Appendix C.4. Figure 12 displays the CRI availability scores.

As can be seen in Figure 12 the scores are all above 70 which is good from a performance standpoint, but it might be an indication that the expected uptimes could be raised. If expected uptimes are consistently being exceeded, some consideration should be given to increase the expected uptimes. This shows that organizations, or at least this particular organization, are getting very good at keeping computer systems available. Therefore, SLAs need to be adjusted to properly reflect the better performance.

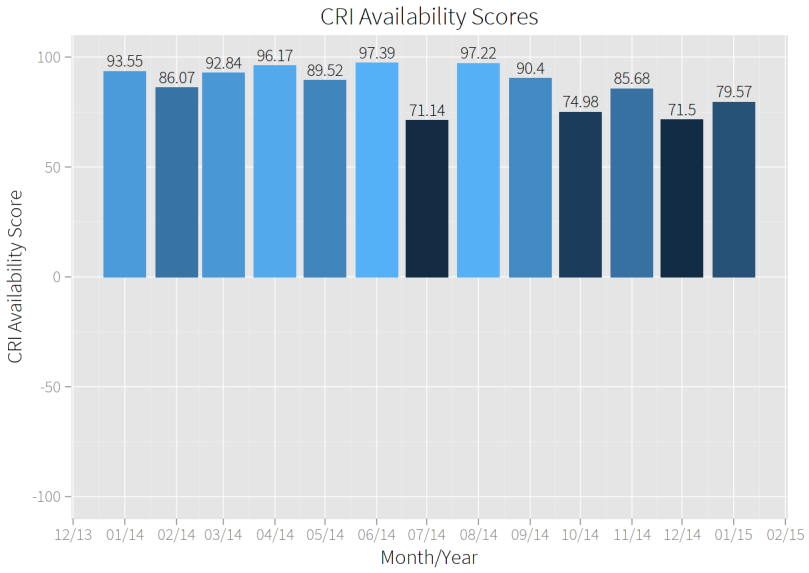

Figure 12: CRI AVAILABILITY SCORES

6.3 SCHEDULE

Figure 19 displays some descriptive statistics for the schedule data. Since the data is mostly dates, the descriptive statistics are limited, so extra values were added to the table after removing four outliers. The new values in the table are: Schedule Duration (Scheduled Finish - Scheduled Start), Difference (Actual Finish - Schedule Finish), and Δ ($\frac{\text{Difference}}{\text{Schedule Duration}}$).

Column	**Min**	**Max**	**Median**	**Mean**
Scheduled Start	2013-06-23	2015-01-24	2014-03-17	2014-04-01
Scheduled Finish	2014-01-31	2015-01-25	2014-08-08	2014-08-02
Actual Start	2013-01-07	2014-10-17	2014-01-15	2014-02-10
Actual Finish	2014-02-20	2014-12-31	2014-08-31	2014-08-11
Schedule Duration	8	480	92	137.7
Difference	-283	149	0.0	-8.919
Δ	-4.75	0.6564	0.0	-0.2641

Column	Min	Max	Median	Mean

41 obs. from 40 Project IDs from mid 2013 to 2015

Table 19: SCHEDULE DATA DESCRIPTIVE STATISTICS

The first task with the schedule data is fitting the data to a distribution. Figure 13 shows a plot of the histogram of the Δs and a Cauchy curve with location = 0.0 and scale = 0.057. The Cauchy distribution appears to be a good fit.

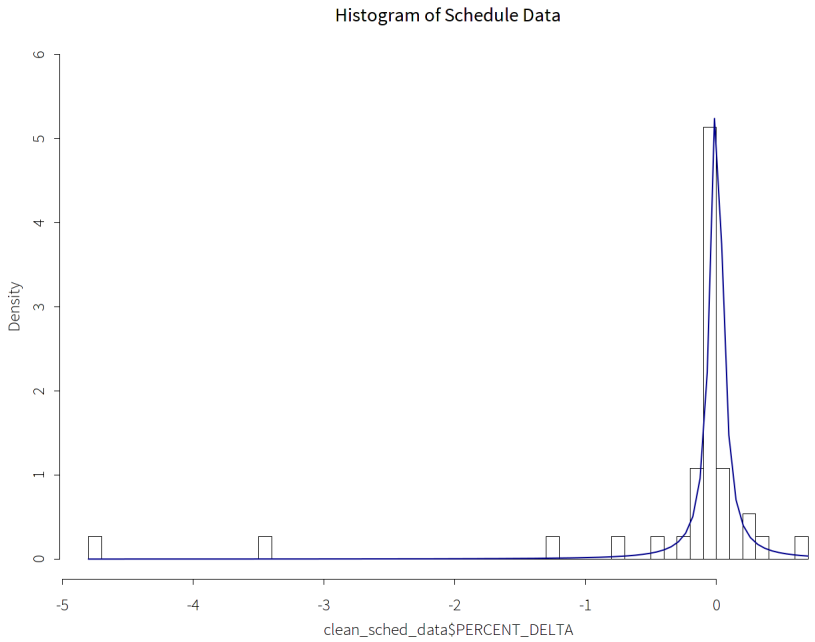

Figure 13: SCHEDULE DATA HISTOGRAM WITH CAUCHY

Now that the distribution has been identified, the next step is determining the CDF for the distribution. For Cauchy, the general CDF is given as follows.

$$CDF(x) = \frac{1}{\pi} \arctan \frac{x - x_0}{\gamma} + \frac{1}{2}$$

where

- x_0 is the location

- γ is the scale

Referencing Equation 4.1.4, the schedule formula for individual project IDs becomes.

$$S_{4_i} = \frac{200}{\pi} \arctan\left(\frac{\Delta_i}{0.057}\right)$$

The CRI schedule scores can be seen in Figure 14. These scores appear more erratic than the other elements. This difference is due to the small number of releases every month.

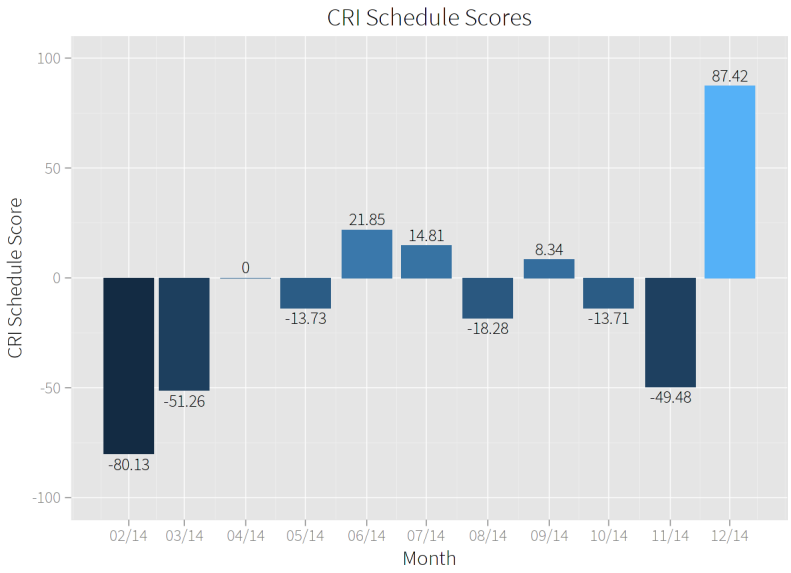

Figure 14: CRI SCHEDULE SCORES

6.4 REQUIREMENTS

The data for requirements are actually counts of story points instead of requirements. That is because story points are a better choice for counting, since they are more uniform in size than raw requirements. Also, the data contained 15 rows of data with a 0 for the number of scheduled requirements. Those 15 rows were removed. Table 20 shows some descriptive statistics for the remaining rows. A histogram for the data can be seen in Appendix C.6.1.

Column	Min	Max	Median	Mean	Variance
Scheduled Requirements	0.5	1248	75.5	136.9	31243.89
Actual Requirements	0	1247	58	123.8	30844.75
$\frac{ActualRequirements}{ScheduledRequirements}$	0	1.2364	1.0	0.8515	0.05

461 obs. from 402 Project IDs from 2010 to 2015

Table 20: REQUIREMENTS DATA DESCRIPTIVE STATISTICS

The multiplier b is set to 1 because none of the historical performance ever exceeded 25% more requirements than scheduled. Thus the upper bound is the number of schedule requirements plus $b = 1$ times the number of scheduled requirements, for an upper bound of twice the scheduled requirements.

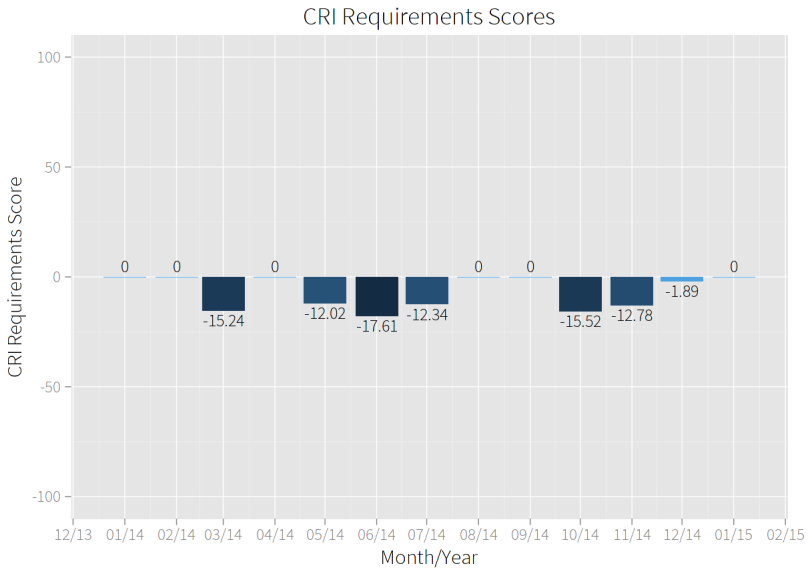

Figure 15: CRI REQUIREMENTS SCORES

Figure 15 displays the CRI requirements scores. The scores are as expected considering the histogram. Many of the projects meet the requirements exactly, and when the requirements are not met, it is usually by a small number. Thus the scores are all 0 or slightly below.

6.5 OVERALL

Now it is time to combine the scores for the overall CRI. Schedule does not have scores for January 2014 or January 2015. Presumably, that is because the SDO within the large financial institution does not schedule projects to be completed in January. Much of the work would need to be completed over Christmas and New Year's Day, but many workers take personal time off of work during that time. Thus some organizations will schedule appropriately. CRI is well suited to handle this problem. Just perform a weighted average over the applicable elements. In this case study, the weights are all equal, so CRI is a straight average.

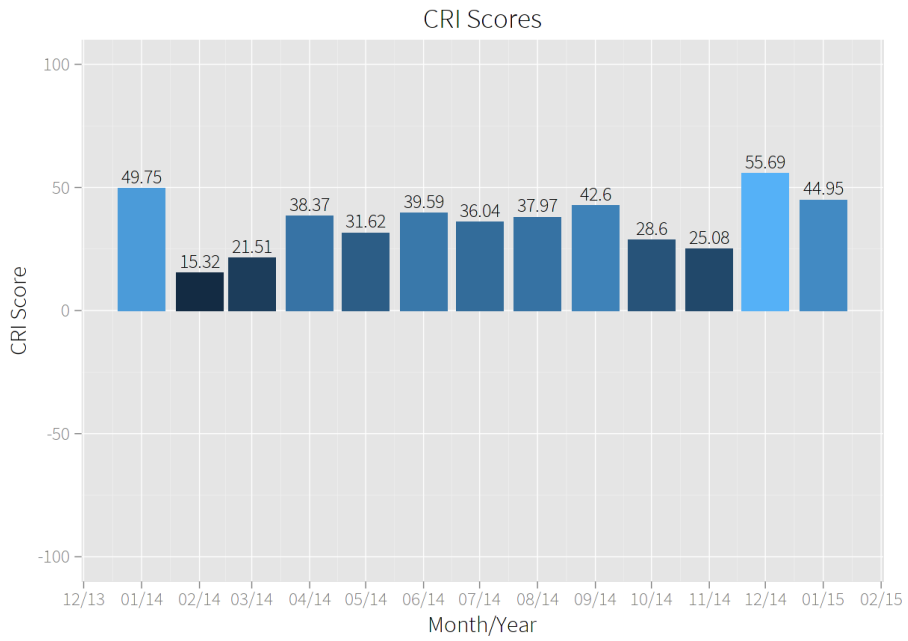

Figure 16: CRI SCORES

Figure 16 displays the CRI scores as computed with the source code from Appendix C.7. The scores are all above 0, so this financial institution has performed consistently better than expected.

6.6 SENSITIVITY AND CORRELATION

To check the sensitivity of the CRI formula, the first method from Section 4.3 is used. Small random values were added to the input values of the formulas. Then the new values were used to calculated the element scores at the respective Application ID or Project ID level. Finally the new CRI element scores were compared with the previous scores from the unaltered data. The goal is to show small changes in the input values result in small changes to the CRI element score. Figure 17 shows histograms of the new scores compared with the previous scores. None of the formulas exhibit too much sensitivity as the histograms all indicate most of the score changes are small, as seen by the high peak of the histograms.

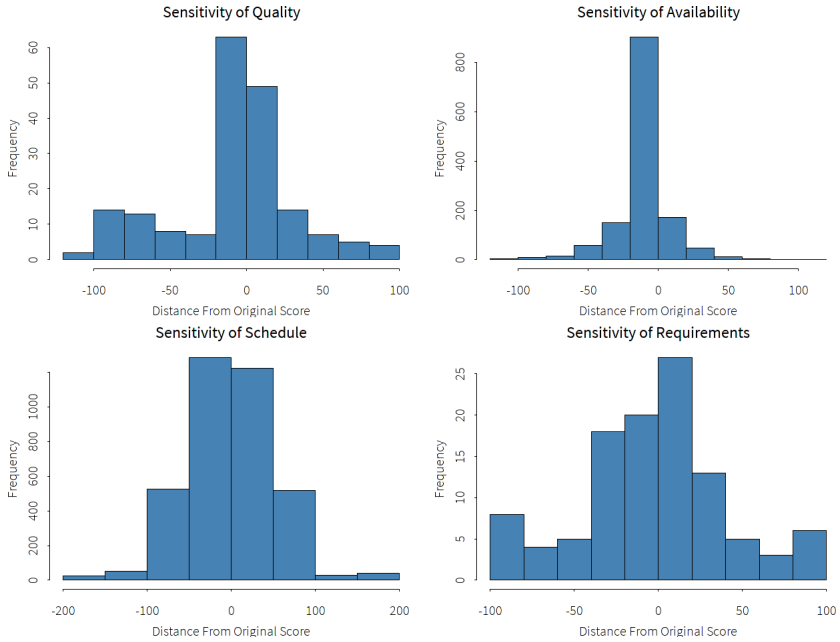

Figure 17: CRI SENSITIVITY ANALYSIS

The random values added were based on the standard deviations of the values in the column. Quality and Requirements used the standard deviation of the matching Application ID or Project ID. This slight difference is due to the wide variation of standard deviations between Application IDs for Quality and Project IDs for Requirements. Table 21 shows the columns that were altered for each element.

CRI Element	Columns Altered
Quality	Development Effort, SIT Defects, UAT Defects
Availability	Percent Uptime
Schedule	$\frac{\text{Schedule Finish - Actual Finish}}{\text{Scheduled Finish - Scheduled Start}}$
Requirements	Actual Requirements Released

Table 21: CRI SENSITIVITY ANALYSIS

The columns were altered by adding random noise from a normal distribution with mean 0 and standard deviation as explained above. The only exception to the normal distribution was for the schedule element. The schedule element drew its noise from a Cauchy distribution with location and scale as defined when determining the distribution of the schedule data. That Cauchy distribution can be seen with the histogram in Figure 13.

Figure 18 shows a scatterplot matrix for the individual element scores. There is no clear upper or downward trend in any of the graphs. Therefore, no correlation appears to exist between any of the elements. The missing element, satisfaction, is the most likely element to exhibit correlation. That is because lower scores on the other four elements will likely result in decreased survey scores, which will result in lower satisfaction scores. Anyhow, the four elements analyzed here remain independent.

7 FUTURE WORK

Due to the number of issues with surveys. One area of future work would be identifying a general set of questions that would best fit CRI. This set would have to include the best number of questions, order of questions and wording of questions. Therefore, any new organization would not have to determine their own survey, but rather just use the predetermined set of questions. It would even be advantageous to build a software system

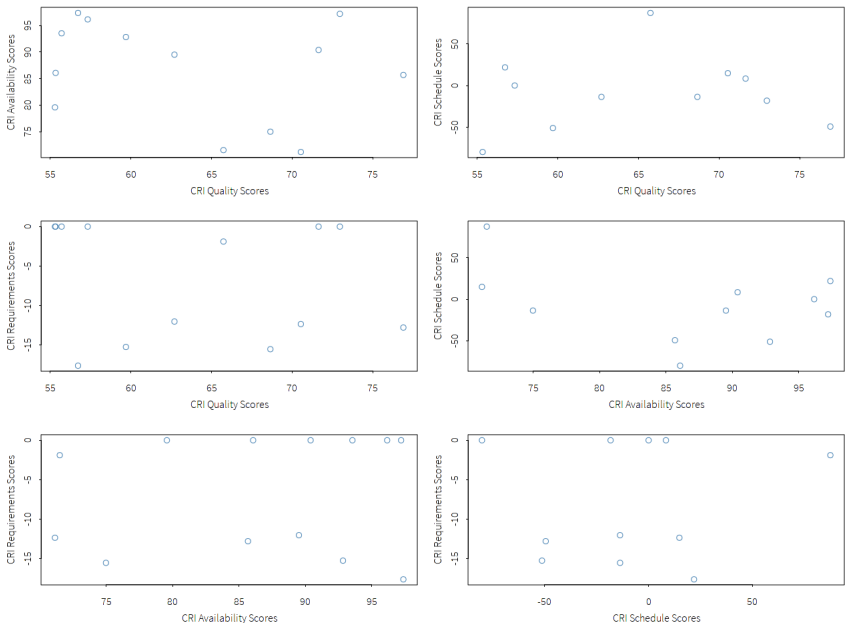

Figure 18: SCATTERPLOT MATRIX OF CRI ELEMENT SCORES

to handle the survey distribution and collection. Ideally, the results would automatically be inserted into the appropriate database tables for CRI scoring.

As shown in Table 16, CRI is currently not addressing two characteristics of the balanced scorecard. If the SDO does not operate on a fixed budget, CRI could be expanded to include another element for financial data. The challenge arises when determining how to relate the SDO performance to finances. The most likely scenario would be a method to either select or predict the expected software sales for a month. Then every month create an CRI element score that reflects how much the expected sales were exceeded or missed. The other missing balanced scorecard characteristic is learning and growth. These are very difficult to quantitatively measure. In this case some possible data points might be: hours of training, number of training courses, number of employees receiving training, number of promotions, or another measure centered around training courses and career growth. Again, once data exists, the problem becomes finding a baseline and measuring with respect to that baseline.

Another area of future work is the expansion of the SDLC-AE to include more

artifacts of the SDLC. The more artifacts and processes that can be collected, the deeper the understanding of the SDLC. All of the data collection combined with better software analytics could lead to true *data-driven software engineering*. Data-driven software engineering is the application of collecting and analyzing historical information about software engineering artifacts in order to accurately predict the outcomes of software engineering projects. This will lead to more informed decisions about software engineering. Figure 19 shows an expansion of the previous SDLC-AE diagram. The new elements are in the unshaded boxes, and they are not exhaustive. Data-driven software engineering should not to be confused with data-driven programming, in which the computer code describes the data instead of the sequence of operations [21].

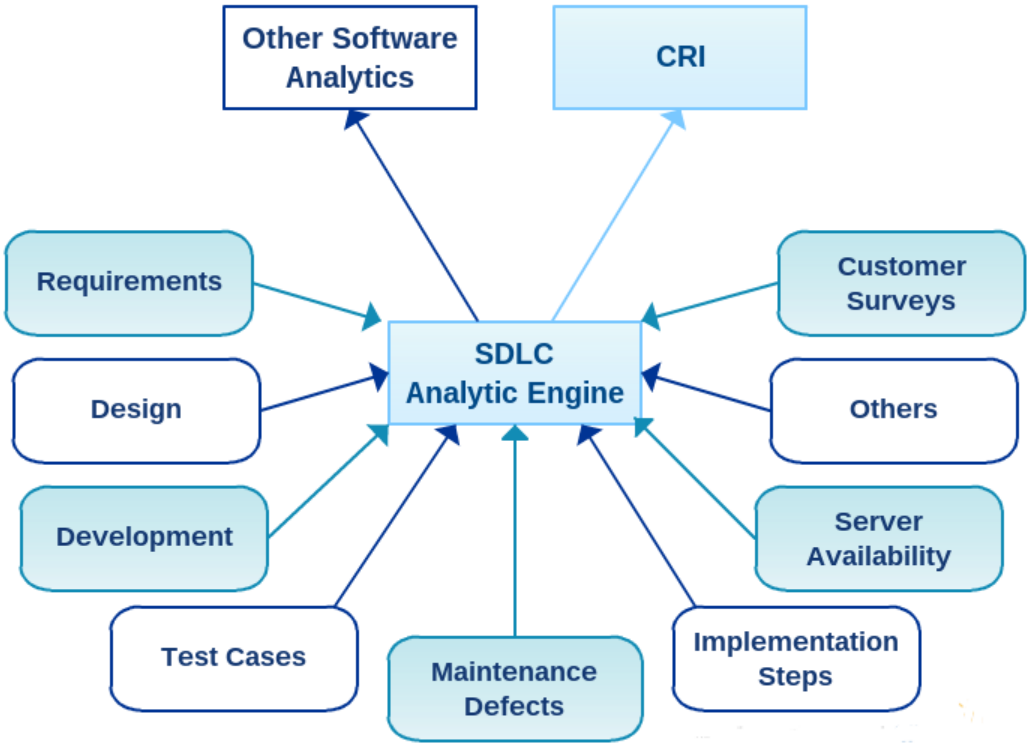

Figure 19: SDLC ANALYTIC ENGINE EXPANSION

Currently, CRI provides a single method to evaluate past performance, but it does not provide any guidance around making more informed future decisions. Some phases are not properly tracked with the initial SDLC-AE and more data can be tracked for the

existing phases. Schedules for each of the individual phases of the SDLC need to be tracked, not just the entire project. Teams need to know how much time is spent in design versus testing. Consideration should also be given to the amount of time required to generate proper testcases. These are just a couple of examples of expansions to the SDLC-AE. These and other advancements could lead to greater insights about SDLC phases that are struggling and need improvement. The SDLC-AE could be expanded to have predictive capabilities.

8 CONCLUSION

There are many metrics that can be used to evaluate an SDO. The entire SDO needs to be measured and analyzed properly, not just the development portion. Knowing which metrics to use and what they all mean can be a daunting task. This dissertation identified the Cumulative Result Indicator (CRI) which is a proposed solution to the difficulty of measuring an SDO by creating a single number score for upper-level management to use to quickly gauge performance.

The following have all been provided in this dissertation.

- CRI defines what five elements of an SDO should be measured. Result indicators are included for quality, availability, satisfaction, schedule, and requirements.
- CRI defines what data needs to be collected to create a score for the five elements.
- CRI defines the formulas to create a score for the five elements.
- A process was outlined to generate the CRI score.
- An SDLC-AE was designed which includes a storage framework for the necessary CRI data.
- A comparison with existing SDO evaluation techniques was presented.

- Finally, an example CRI score was calculated with data provided by an SDO organization within a large financial institution.

Now SDOs have a technique to consistently measure performance over time. CRI will help upper-level management identify the areas of the organization that need attention and those that do not. CRI will save time required to evaluate performance since the result is a single number.

APPENDIX

A DETAILED STEPS OF THE SDLC

The SDLC often contains more than the five basic steps of requirements, design, implementation, testing, and deployment/maintenance. Those are the high-level phases, but many steps are required to complete each phase. The following list provides a more detailed list of what needs to be accomplished in the entire life cycle of software development. These steps do not need to occur in a sequential fashion.

- Identify the Work/Task/Project
  - Get Initial Idea
  - Obtain Details
- Estimate
  - Create an Estimate (What is included? What is the output? days/dollars/hours/reqs)
  - Obtain Approval
  - Quit or Go Forward
- Document Requirements
  - Identify the Requirements
  - Detail the Requirements
- Design The Software
  - Find System Integrations
  - Identify Functional Specs
  - Detail the Functional Specs
- Development of all the tasks in Design and Requirements

  - Identify the Coding Tasks
  - Write the Code/Develop the solution
  - Write the Unit Tests
- Test
  - Create Test Plans and Cases
  - Run Test Plans and Cases
- Deployment
  - Create Deployment Steps
  - Run Deployment Steps
- Maintenance
  - Capture Bugs
  - Survey Users
- Start Again

B SDLC-AE SOURCE CODE

B.1 SQL CODE - DATA TABLES

```
-- create the raw data tables

CREATE TABLE QUALITY_RAW (
    QUALITY_RAW_ID RAW(16) NOT NULL PRIMARY KEY,
    APPLICATION_ID VARCHAR2(64) NOT NULL,
    FREQ_DATE DATE NOT NULL,
    DEV_EFFORT NUMBER(10,0) NOT NULL,
    TEST_EFFORT NUMBER(10,0),
    SIT_DEFECTS NUMBER(10,0),
    UAT_DEFECTS NUMBER(10,0),
    PROD_DEFECTS NUMBER(10,0) NOT NULL,
    INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
    UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
);

CREATE TABLE AVAILABILITY_RAW (
```

```
17      AVAILABILITY_RAW_ID RAW(16) NOT NULL PRIMARY KEY,
18      SERVICE_ID VARCHAR2(64) NOT NULL,
19      FREQ_DATE DATE NOT NULL,
20      UPTIME NUMBER(10,5),
21      SCHED_DOWNTIME NUMBER(10,5),
22      UNSCHED_DOWNTIME NUMBER(10,5),
23      PERCENT_UPTIME NUMBER(10,5) NOT NULL,
24      EXPECT_PERCENT_UPTIME NUMBER(10,5) NOT NULL,
25      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
26      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
27  );
28
29  CREATE TABLE SATISFACTION_RAW (
30      SATISFACTION_RAW_ID RAW(16) NOT NULL PRIMARY KEY,
31      QUESTION_ID VARCHAR2(64) NOT NULL,
32      FREQ_DATE DATE NOT NULL,
33      QUESTION_TEXT VARCHAR2(1024),
34      RESPONDENT_ID VARCHAR2(128),
35      RESPONSE NUMBER(5,0) NOT NULL,
36      RESPONSE_DATE DATE,
37      APPLICATION_ID VARCHAR2(64),
38      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
39      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
40  );
41
42  CREATE TABLE SCHEDULE_RAW (
43      SCHEDULE_RAW_ID RAW(16) NOT NULL PRIMARY KEY,
44      PROJECT_ID VARCHAR2(64),
45      FREQ_DATE DATE NOT NULL,
46      SCHED_START_DATE DATE NOT NULL,
47      SCHED_FINISH_DATE DATE NOT NULL,
48      ACTUAL_START_DATE DATE,
49      ACTUAL_FINISH_DATE DATE NOT NULL,
50      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
51      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
52  );
53
54  CREATE TABLE REQUIREMENTS_RAW (
55      REQUIREMENTS_RAW_ID RAW(16) NOT NULL PRIMARY KEY,
56      PROJECT_ID VARCHAR2(64),
57      FREQ_DATE DATE NOT NULL,
58      SCHED_REQ NUMBER(10,0) NOT NULL,
59      ACTUAL_REQ NUMBER(10,0) NOT NULL,
60      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
61      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
62  );
63
```

```
-- remove the raw tables
--DROP TABLE QUALITY_RAW;
--DROP TABLE AVAILABILITY_RAW;
--DROP TABLE SATISFACTION_RAW;
--DROP TABLE SCHEDULE_RAW;
--DROP TABLE REQUIREMENTS_RAW;
```

B.2 SQL CODE - SCORE TABLES

```
-- Create the scoring tables

CREATE TABLE QUALITY_SCORE (
    QUALITY_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
    APPLICATION_ID VARCHAR2(64) NOT NULL,
    FREQ_DATE DATE NOT NULL,
    SCORE NUMBER(10,5) NOT NULL,
    ACTIVE CHAR DEFAULT 'Y' NOT NULL,
    INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
    UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
);

CREATE TABLE AVAILABILITY_SCORE (
    AVAILABILITY_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
    SERVICE_ID VARCHAR2(64) NOT NULL,
    FREQ_DATE DATE NOT NULL,
    SCORE NUMBER(10,5) NOT NULL,
    ACTIVE CHAR DEFAULT 'Y' NOT NULL,
    INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
    UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
);

CREATE TABLE SATISFACTION_SCORE (
    AVAILABILITY_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
    QUESTION_ID VARCHAR2(64) NOT NULL,
    FREQ_DATE DATE NOT NULL,
    SCORE NUMBER(10,5) NOT NULL,
    ACTIVE CHAR DEFAULT 'Y' NOT NULL,
    INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
    UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
);

CREATE TABLE SCHEDULE_SCORE (
    QUALITY_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
    PROJECT_ID VARCHAR2(64) NOT NULL,
    FREQ_DATE DATE NOT NULL,
    SCORE NUMBER(10,5) NOT NULL,
    ACTIVE CHAR DEFAULT 'Y' NOT NULL,
```

```
39      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
40      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
41  );
42
43  CREATE TABLE REQUIREMENTS_SCORE (
44      REQUIREMENTS_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
45      PROJECT_ID VARCHAR2(64) NOT NULL,
46      FREQ_DATE DATE NOT NULL,
47      SCORE NUMBER(10,5) NOT NULL,
48      ACTIVE CHAR DEFAULT 'Y' NOT NULL,
49      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
50      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
51  );
52
53  -- remove the scoring tables
54  --DROP TABLE QUALITY_SCORE;
55  --DROP TABLE AVAILABILITY_SCORE;
56  --DROP TABLE SATISFACTION_SCORE;
57  --DROP TABLE SCHEDULE_SCORE;
58  --DROP TABLE REQUIREMENTS_SCORE;
```

B.3 SQL CODE - FINAL SCORE TABLES

```
1   -- create the ELEMENT table
2
3   CREATE TABLE ELEMENT (
4       ELEMENT_ID NUMBER(10,0) NOT NULL PRIMARY KEY,
5       NAME VARCHAR2(64) NOT NULL,
6       DESCRIPTION VARCHAR(255),
7       INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
8       UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
9   );
10
11  -- Add the CRI score types to the table
12  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
13      VALUES (1,'QUALITY');
14  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
15      VALUES (2,'AVAILABILITY');
16  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
17      VALUES (3,'SATISFACTION');
18  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
19      VALUES (4,'SCHEDULE');
20  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
21      VALUES (5,'REQUIREMENTS');
22  INSERT INTO ELEMENT (ELEMENT_ID,NAME)
23      VALUES (6,'OVERALL');
24
```

```
25
26 -- create the overall score table
27 CREATE TABLE CRI_SCORE (
28      CRI_SCORE_ID RAW(16) NOT NULL PRIMARY KEY,
29      ELEMENT_ID NUMBER(10,0) NOT NULL REFERENCES ELEMENT(
       ELEMENT_ID),
30      FREQ_DATE DATE NOT NULL,
31      SCORE NUMBER(10,5) NOT NULL,
32      ACTIVE CHAR DEFAULT 'Y' NOT NULL,
33      INSERT_DATE DATE DEFAULT SYSDATE NOT NULL,
34      UPDATE_DATE DATE DEFAULT SYSDATE NOT NULL
35 );
36
37 -- remove ELEMENT table
38 --DROP TABLE CRI_SCORE;
39 --DROP TABLE ELEMENT;
```

C CASE STUDY SOURCE CODE

A full set of the source code and applicable output is available at [81].

C.1 QUALITY HISTORICAL R CODE AND ANALYSIS

```
1  ### Load Raw Quality Data
2  setAs("character","myDate",
3        function(from) {as.Date(from, format="%m/%d/%Y")} )
4  setClass('myDate')
5
6  quality_raw <- read.csv('data/quality_raw.csv',
7                          colClasses=c('factor',
8                                       'myDate',
9                                       'numeric',
10                                      'numeric',
11                                      'numeric',
12                                      'numeric') )
13
14 ### Get descriptive statistics
15 str(quality_raw)
16 summary(quality_raw)
17 var(quality_raw)
18
19 ### Find the Baseline Quality Function
20 #### Use data prior to 2014
21 history_quality_raw =
22         quality_raw[quality_raw$MONTH_DT
23                     <= as.Date('2013-12-31') ,]
```

```
24
25 #### Create some plots of the historical quality data
26 par(mfrow=c(2,2))
27 plot(history_quality_raw$DEV_EFF,
28      history_quality_raw$PROD_DFTS,
29       xlab='DEV_EFF', ylab='PROD_DFTS', col='steelblue')
30 plot(history_quality_raw$SIT_DFTS,
31      history_quality_raw$PROD_DFTS,
32       xlab='SIT_DFTS', ylab='PROD_DFTS', col='steelblue')
33 plot(history_quality_raw$DEV_EFF,
34      history_quality_raw$PROD_DFTS,
35       xlab='UAT_DFTS', ylab='PROD_DFTS', col='steelblue')
36
37
38 ##### Remove the outlier data point with 1216 PROD_DFTS
39 history_quality_clean = history_quality_raw[history_quality_raw$
     PROD_DFTS < 1000,]
40
41 ##### create the model after dropping the
42 ##### data point with over 1000 PROD_DFTS
43 baseline_quality_function = lm(PROD_DFTS ~ DEV_EFF + SIT_DFTS +
     UAT_DFTS,
44             data=history_quality_clean )
45 summary(baseline_quality_function)
46
47 par(mfrow=c(1,2))
48 qqnorm(baseline_quality_function$resid,col='steelblue')
49 qqline(baseline_quality_function$resid,col='steelblue')
50 summary(baseline_quality_function)$sigma
51 plot(baseline_quality_function$fitted,abs(baseline_quality_
     function$resid),
52       col='steelblue',
53      main='Fitted vs Resid',
54      xlab='Fitted',
55      ylab='Absolute Value of Residuals')
56 pairs(history_quality_clean[ ,c('DEV_EFF','SIT_DFTS','UAT_DFTS')
     ],col='steelblue')
57
58 #### source code for ridge regression
59 #### did not yield better results
60 install.packages('ridge')
61 library('ridge')
62 rd = linearRidge(PROD_DFTS ~ DEV_EFF
63             + SIT_DFTS + UAT_DFTS ,
64             data=history_quality_clean, nPCs=1)
65 summary(rd)
```

The source code for the selected baseline quality function can be seen above. The output for the linear model can be seen below.

```
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  5.917433    2.959339    2.000    0.0459 *
DEV_EFF      0.034942    0.001773   19.709   < 2e-16 ***
SIT_DFTS    -0.362998    0.048068   -7.552  1.18e-13 ***
UAT_DFTS     1.048225    0.143276    7.316  6.25e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77.43 on 794 degrees of freedom
Multiple R-squared:  0.7188,    Adjusted R-squared:  0.7178
F-statistic: 676.6 on 3 and 794 DF,  p-value: < 2.2e-16
```

Some diagnostic plots for the baseline quality function can be seen in Figure 20. The normal probability plot, a.k.a. Q-Q Plot, shows the errors are not exactly normally distributed, but the baseline quality function had good predictive power as shown by the high R^2. The fitted versus residuals plot indicates a lack of heteroscedasticity.

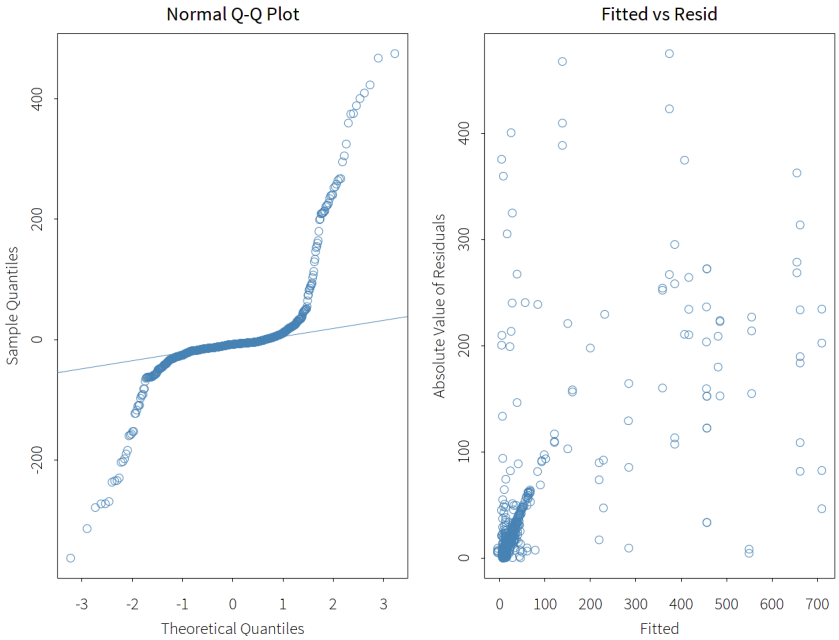

Figure 20: QUALITY DIAGNOSTIC PLOTS

Figure 21 shows the presence of some possible multicolinearity. As a result, ridge regression was used to create a model, but the results were very similar to the original

baseline quality function. Therefore, ridge regression was not chosen.

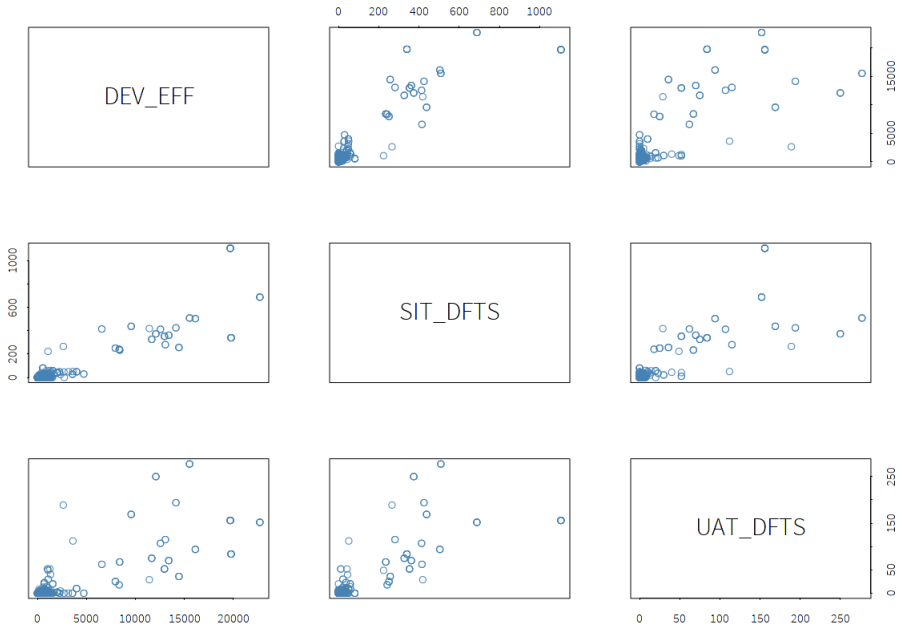

Figure 21: QUALITY PAIRS PLOT OF INDEPENDENT VARIABLES

C.2 BAR CHART - R CODE

This is a code for generating a bar chart. It is used for reporting the scores of all the elements.

```
library('ggplot2')
install.packages('zoo')
library('zoo')
library('scales')

barchart <- function(data, main='CRI Scores', xlab='Month/Year',
    ylab='CRI Score') {

     ggplot(data, aes(x=as.Date(MONTH_DT), y=SCORE, fill=SCORE)) +
       geom_bar(stat = "identity") +
       geom_text(aes(label=round(SCORE, 2), vjust=ifelse(sign(
    SCORE)>=0, -.4, 1.4)), size=9 ) +
       guides(fill=FALSE) +
       ylim(-100,100) +
```

```
        scale_x_date(labels = date_format("%m/%y"), breaks = date_
    breaks("month")) +
        ggtitle(main) +
        ylab(ylab) +
        xlab(xlab)
}
```

C.3 QUALITY SCORES - R CODE

```
### Now calculate the Quality scores
#### Get data for 2014 and newer
current_quality =
        quality_raw[quality_raw$MONTH_DT
                    > as.Date('2013-12-31') ,]

current_quality$PREDICTION =
        predict(baseline_quality_function, newdata=current_
    quality)
current_quality$SCORE = 100 *
        ifelse(current_quality$PREDICTION
               >= current_quality$PROD_DFTS,
                  (current_quality$PREDICTION
                    - current_quality$PROD_DFTS)
                    /current_quality$PREDICTION,
                  (current_quality$PREDICTION
                    - current_quality$PROD_DFTS)
                    /(summary(baseline_quality_function)$sigma^2)
                                        )

qual_scores = aggregate(SCORE ~ MONTH_DT, current_quality, mean )
qual_scores$MONTH_DT = as.yearmon(qual_scores$MONTH_DT, "%Y-%B")
qual_scores$SCORE = round(qual_scores$SCORE, 2)
qual_scores

barchart(qual_scores, main='CRI Quality Scores', ylab='CRI
    Quality Score',xlab='Month/Year')
```

C.4 AVAILABILITY SCORES - R CODE

```
library('sense')
source('barchart.R')

# Load Data
# ---------
#
# Load the Availability data.
```

```
setAs("character","myDate",
      function(from) {as.Date(from, format="%m/%d/%Y")} )
setClass('myDate')

avail_raw <- read.csv('data/availability_raw.csv',
                      colClasses=c('factor',
                                   'myDate',
                                   'numeric',
                                   'numeric'))

summary(avail_raw)
str(avail_raw)

# trim to only 2014 and newer
avail_data =
        avail_data[avail_data$SLA_DATE >= as.Date('2014-01-01')
    ,]

avail_data$SCORE =  ifelse(
  avail_data$ACTUAL <= avail_data$EXPECTED,
  (avail_data$ACTUAL - avail_data$EXPECTED)/avail_data$EXPECTED,
  (avail_data$ACTUAL - avail_data$EXPECTED)/(1 - avail_data$
   EXPECTED)
  )

avail_scores = aggregate(SCORE ~ SLA_DATE, avail_data, mean )
avail_scores$SLA_DATE = as.yearmon(avail_scores$SLA_DATE, "%Y-%B")
    )
avail_scores$SCORE = round(100* avail_scores$SCORE, 2)

barchart(avail_scores, main='CRI Availability Scores',
         ylab='CRI Quality Score',xlab='Month/Year')
```

C.5 SCHEDULE SCORES - R CODE

```
source('barchart.R')

## Load the Schedule data.
setAs("character","myDate", function(from) {as.Date(from, format=
    "%m/%d/%Y")} )
setClass('myDate')

schedule_raw <- read.csv('data/schedule_raw.csv',
                      colClasses=c('factor',
                                   'myDate',
                                   'myDate',
                                   'myDate',
```

```
12                                          'myDate',
13                                          'myDate'))
14 summary(schedule_raw)
15 str(schedule_raw)
16
17 #### Clean up data by removing all rows with
18 #### schedule start, schedule end, and actual end
19 clean_sched_data = schedule_raw[ !is.na(schedule_raw$ACTUAL_
   FINISH) & !is.na(schedule_raw$SCHED_START) & !is.na(schedule_
   raw$SCHED_FINISH),]
20
21 #### Find the estimated duration,
22 #### and the delta from the estimated finish
23 #### and percent
24 clean_sched_data$EST_DUR = as.numeric(clean_sched_data$SCHED_
   FINISH - clean_sched_data$SCHED_START+1)
25 clean_sched_data$DELTA = as.numeric(clean_sched_data$SCHED_FINISH
    - clean_sched_data$ACTUAL_FINISH )
26 clean_sched_data$PERCENT_DELTA = clean_sched_data$DELTA/clean_
   sched_data$EST_DUR
27
28 #### Remove some of the outlier data
29 clean_sched_data = clean_sched_data[abs(clean_sched_data$PERCENT_
   DELTA) < 20 , ]
30 summary(clean_sched_data)
31 str(clean_sched_data)
32
33 ### Fit a Distribution to the data
34
35 location = median(clean_sched_data$PERCENT_DELTA)
36 scale = IQR(clean_sched_data$PERCENT_DELTA)/2
37 c(location, scale)
38 h = hist(clean_sched_data$PERCENT_DELTA, breaks = 70, prob=TRUE,
   ylim=c(0,6), main='Histogram of Schedule Data')
39 x=seq(-5,5,length=200)
40 curve(dcauchy(x, location, scale ),
41           col="darkblue", lwd=2, add=TRUE)
42
43 ### Now find the CRI scores
44 sched_data = clean_sched_data
45 sched_data$SCORE = (200/pi)*atan(sched_data$PERCENT_DELTA /scale)
46 sched_data$MONTH_DT = as.yearmon(sched_data$ACTUAL_FINISH, "%Y-%B
   ")
47 sched_scores = aggregate(SCORE ~ MONTH_DT, sched_data, mean )
48 sched_scores
49
```

```
50 barchart(sched_scores, main='CRI Schedule Scores', ylab='CRI
     Schedule Score')
```

C.6 REQUIREMENTS SCORES - R CODE

C.6.1 REQUIREMENTS HISTOGRAM

Figure 22 shows a histogram of the data for requirements. The histogram is based upon the fraction, $\frac{ScheduledRequirements}{ActualRequirements}$. The values where the fraction equals one have been left out of the histogram. So, the histogram show only the data that did not deliver exactly on the number of requirements. Rarely are more requirements actually delivered than what was scheduled. On the opposite side, the histogram bars shrink as they approach zero indicating that it is more common to miss a few requirements than all the requirements.

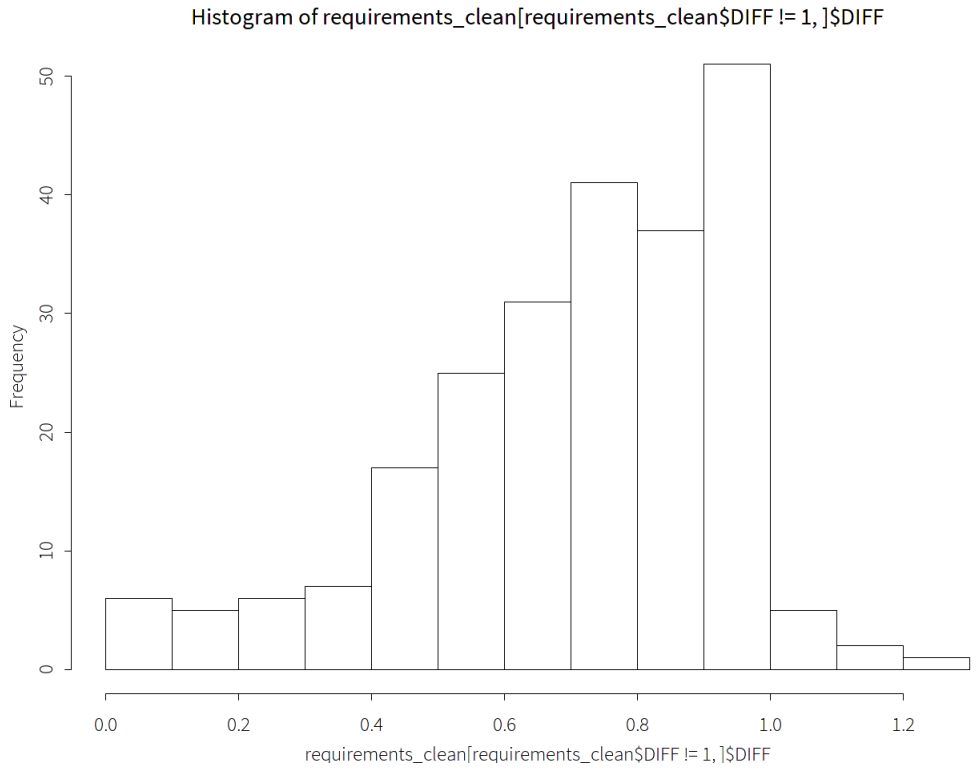

Figure 22: REQUIREMENTS DATA HISTOGRAM (ACTUAL/SCHEDULED)

C.6.2 REQUIREMENTS R CODE

```
1 library('sense')
```

```
source('barchart.R')

### Now read in the requirements data
requirements_raw = read.csv('data/requirements_raw.csv',
                            colClasses=c('factor',
                                         'Date',
                                         'numeric',
                                         'numeric') )

##### remove rows where COMPLETED and SCHEDULED are both 0
requirements_clean = requirements_raw[requirements_raw$SCHEDULED
    != 0, ]

requirements_clean$DIFF = requirements_clean$COMPLETED/
    requirements_clean$SCHEDULED
summary(requirements_clean)
str(requirements_clean)
var(requirements_clean$SCHEDULED)
var(requirements_clean$COMPLETED)
var(requirements_clean$DIFF)

hist(requirements_clean[requirements_clean$DIFF != 1,]$DIFF)

##### Now calculate the requirements CRI score
##### but only for year 2014 and 2015
current_requirements = requirements_clean[requirements_clean$
    MONTH_DT
                        > as.Date('2013-12-31'), ]
current_requirements$MONTH_DT = as.yearmon(as.Date(current_
    requirements$MONTH_DT), "%Y-%B")
current_requirements$SCORE = 100*(current_requirements$COMPLETED-
    current_requirements$SCHEDULED)/current_requirements$SCHEDULED

req_scores = aggregate(SCORE ~ MONTH_DT, current_requirements,
    mean )
req_scores$SCORE = round(req_scores$SCORE, 2)
req_scores

barchart(req_scores, main='CRI Requirements Scores', ylab='CRI
    Requirements Score',xlab='Month/Year')
```

C.7 OVERALL SCORES - R CODE

```

### Combine the scores
# overall

```

```
# overall

full = merge(x = qual_scores, y = avail_scores, by = "MONTH_DT",
    all = TRUE, suffixes=c('_QUAL','_AVAIL'))
full = merge(x = full, y = sched_scores, by = "MONTH_DT", all =
    TRUE, suffixes='_SCHED' )
colnames(full)[4] = 'SCORE_SCHED'
full = merge(x = full, y = req_scores, by = "MONTH_DT", all =
    TRUE, suffixes=c('','_REQ') )
colnames(full)[5] = 'SCORE_REQ'

full$SCORE = apply(full[,seq(2,5)], 1,
                   function(x) {sum(x,na.rm=TRUE)/sum(!is.na(x))}
    )

barchart(full)
```

D ADDITIONAL SDLC DATA NEEDS

This appendix describes some additional SDLC attributes that could be tracked to help improve estimation. The SDLC-AE could be expanded to include the following.

D.1 ESTIMATION

The following are additional data points that could be tracked for project estimation.

- Change to database structure
- Modify database data
- Create a new database, number of new databases
- Server configuration changes required
- New servers required
- Number of people involved
- Number of (sub)systems involved
- Estimation date
- Number of days allowed

- List of other attributes
- Number of screens involved
- Actual values (hours, days, dollars)
- Estimated values
  - Estimated development hours: The number of development hours estimated for a project, this is just developer hours
  - Estimated documentation hours: The number of documentation hours estimated for a project
  - Estimated testing hours: The number of testing hours estimated for a project
  - Estimated deployment hours: The number of estimated hours required to deploy the project

D.2 REQUIREMENTS

The following are additional data points that could be tracked for the requirements phase of an SDLC project.

- Title
- Description
- Author
- Project
- Date
- Comments
  - Date
  - Comment text
  - Author

D.3 DEVELOPMENT

The following are additional data points that could be tracked for the development phase of an SDLC project.

- Project
- Release
- List of files
- Author
- Date started
- Completion date
- Number of unit tests
- Lines of code
- Percentage of automated test coverage
- Others

D.4 TESTING

The following are additional data points that could be tracked during, before, and after the testing phase of an SDLC project.

- Project
- Release
- Title
- description]
- Author]
- Date started
- Date executed
- Status (pending, pass, fail)

- Comments
  - Date
  - Comment text
  - Author

D.5 IMPLEMENTATION

The following are additional data points that could be tracked for the implementation of a project.

- Project
- Release
- date Entered
- Date Scheduled
- Date Executed
- Ordering/Prerequisites
- Comments
  - date
  - Comment text
  - Author

D.6 MAINTENANCE (DEFECTS)

The following are additional data points that could be tracked for maintenance of a project.

- Project
- Release
- Description

- Date Entered
- Date fixed
- Comments
  - Date
  - Comment text
  - Author

REFERENCES

[1] M. Andreessen, "Why software is eating the world," *The Wall Street Journal*, Aug. 2011. [Online]. Available: http://goo.gl/MXk4TS.

[2] Ž. Antolić, "An example of using key performance indicators for software development process efficiency evaluation," in *MIPRO 2008: 31st International Convention on Information and Communication Technology, Electronics and Microelectronics, May 26-30, 2008, Opatija Croatia. Microelectronics, electronics and electronic technologies, MEET.. Grid and visualization systems, GVS*, P. Biljanović, K. Skala, and G. . V. Systems, Eds., vol. 1, MIPRO, 2008, ISBN: 9789532330366.

[3] S. Ashmore and K. Runyan, *Introduction to Agile Methods*, 1st. Addison-Wesley Professional, 2014, ISBN: 032192956X, 9780321929563.

[4] B. Aulet, "Disciplined entrepreneurship: 24 steps to a successful startup," in. 2013, ch. STEP 19: Calculate the Cost of Customer Acquisition (COCA), ISBN: 1-118692-28-4.

[5] N. Ayewah, W. Pugh, J. D. Morgenthaler, J. Penix, and Y. Zhou, "Evaluating static analysis defect warnings on production software," in *Proceedings of the 7th ACM SIGPLAN-SIGSOFT Workshop on Program Analysis for Software Tools and Engineering*, ser. PASTE '07, New York, NY, USA: ACM, 2007, pp. 1–8, ISBN: 978-1-59593-595-3. DOI: 10.1145/1251535.1251536.

[6] K. Beck, M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R. C. Martin, S. Mellor, K. Schwaber, J. Sutherland, and D. Thomas. (2001). Manifesto for agile software development, [Online]. Available: http://www.agilemanifesto.org/.

[7] H. D. Benington, "Production of large computer programs," in *Proceedings of the 9th International Conference on Software Engineering*, ser. ICSE '87, Los Alamitos, CA, USA: IEEE Computer Society Press, 1987, pp. 299–310, ISBN: 0-89791-216-0.

[8] B. W. Boehm, *Software Engineering Economics*, 1st. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1981, ISBN: 0138221227.

[9] B. W. Boehm, "A spiral model of software development and enhancement," *SIGSOFT Software Engineering Notes*, vol. 11, no. 4, pp. 14–24, Aug. 1986, ISSN: 0163-5948. DOI: 10.1145/12944.12948.

[10] B. W. Boehm, "A spiral model of software development and enhancement," *Computer*, vol. 21, no. 5, pp. 61–72, May 1988, ISSN: 0018-9162. DOI: 10.1109/2.59.

[11] B. W. Boehm and V. R. Basili, "Software defect reduction top 10 list," *Computer*, vol. 34, no. 1, pp. 135–137, Jan. 2001, ISSN: 0018-9162. DOI: 10.1109/2.962984.

[12] B. W. Boehm and W. J. Hansen, "Spiral development: Experience, principles, and refinements," Carnegie Mellon Software Engineering Institute, Tech. Rep., Jul. 2000, Special Report CMU/SEI-2000-SR-008.

[13] R. P. Buse and T. Zimmermann, "Analytics for software development," in *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research*, ser. FoSER '10, New York, NY, USA: ACM, 2010, pp. 77–80, ISBN: 978-1-4503-0427-6. DOI: 10.1145/1882362.1882379. [Online]. Available: http://research.microsoft.com/pubs/136301/MSR-TR-2010-111.pdf.

[14] R. P. Buse and T. Zimmermann, "Information needs for software development analytics," in *Proceedings of the 34th International Conference on Software Engineering*, ser. ICSE '12, Piscataway, NJ, USA: IEEE Press, 2012, pp. 987–996, ISBN: 978-1-4673-1067-3.

[15] R. Charette, "Why software fails [software failure]," *IEEE Spectrum*, vol. 42, no. 9, pp. 42–49, Sep. 2005, ISSN: 0018-9235. DOI: 10.1109/MSPEC.2005.1502528.

[16] CMMI Product Team, "Cmmi for development, version 1.3," Carnegie Mellon Software Engineering Institute (SEI), http://goo.gl/MBESq0, Tech. Rep., Nov. 2010.

[17] T. Copeland, *PMD Applied: An Easy-to-use Guide for Developers*, ser. An easy-to-use guide for developers. Centennial Books, 2005, ISBN: 9780976221418.

[18] E. Cowles and E. Nelson, *An Introduction to Survey Research*, 1st. New York, NY, USA: Business Expert Press, Jan. 2015, ISBN: 978-1-60649-818-7.

[19] J. Czerwonka, N. Nagappan, W. Schulte, and B. Murphy, "Codemine: Building a software development data analytics platform at microsoft," *IEEE Software*, vol. 30, no. 4, pp. 64–71, 2013, ISSN: 0740-7459. DOI: 10.1109/MS.2013.68.

[20] A. Damodaran. (Feb. 2015). Statistical distributions, New York University, [Online]. Available: http://goo.gl/Jp7mtn.

[21] (Mar. 2015). Data-driven programming, Wikipedia, [Online]. Available: http://en.wikipedia.org/wiki/Data-driven_programming.

[22] T. DeMarco, "Software engineering: An idea whose time has come and gone?" *IEEE Software*, vol. 26, no. 4, pp. 96–96, Jul. 2009, ISSN: 0740-7459. DOI: 10.1109/MS.2009.101.

[23] A. B. Downey, *Think Stats: Probability and Statistics for Programmers*. O'Reilly Media, 2011, p. 138. [Online]. Available: http://greenteapress.com/thinkstats/.

[24] D. J. Dubois and G. Tamburrelli, "Understanding gamification mechanisms for software development," in *Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering*, ACM, 2013, pp. 659–662.

[25] C. Ebert, T. Liedtke, and E. Baisch, "Improving reliability of large software systems," *Annals of Software Engineering*, vol. 8, no. 1-4, pp. 3–51, Aug. 1999, ISSN: 1022-7091. DOI: 10.1023/A:1018971212809.

[26] K. E. Emam and A. G. Koru, "A replicated survey of it software project failures," *IEEE Software*, vol. 25, no. 5, pp. 84–90, Sep. 2008, ISSN: 0740-7459. DOI: 10.1109/MS.2008.107.

[27] M. Faizan, M. N. A. Khan, and S. Ulhaq, "Contemporary trends in defect prevention: A survey report," *International Journal of Modern Education and Computer Science (IJMECS)*, vol. 4, no. 3, p. 14, 2012.

[28] W. A. Florac and A. D. Carleton, *Measuring the Software Process*. Boston: Addison Wesley, 1999.

[29] C. Giardino, M. Unterkalmsteiner, N. Paternoster, T. Gorschek, and P. Abrahamsson, "What do we know about software development in startups?" *IEEE Software*, vol. 31, no. 5, pp. 28–32, Sep. 2014, ISSN: 0740-7459. DOI: 10.1109/MS.2014.129.

[30] S. Godfrey. (Oct. 2011). Characteristics of capability maturity model, [Online]. Available: http://goo.gl/MpNx9b.

[31] A. L. Goel and M. Shin, "Software engineering data analysis techniques (tutorial)," in *Proceedings of the 19th International Conference on Software Engineering*, ser. ICSE '97, New York, NY, USA: ACM, 1997, pp. 667–668, ISBN: 0-89791-914-9. DOI: 10.1145/253228.253816.

[32] M. Halkidi, D. Spinellis, G. Tsatsaronis, and M. Vazirgiannis, "Data mining in software engineering," *Intelligent Data Analysis*, vol. 15, no. 3, pp. 413–441, Aug. 2011, ISSN: 1088-467X.

[33] M. H. Halstead, *Elements of Software Science (Operating and Programming Systems Series)*. New York, NY, USA: Elsevier Science Inc., 1977, ISBN: 0444002057.

[34] A. E. Hassan, A. Hindle, P. Runeson, M. Shepperd, P. Devanbu, and S. Kim, "Roundtable: What's next in software analytics," *IEEE Software*, vol. 30, no. 4, pp. 53–56, Jul. 2013, ISSN: 0740-7459. DOI: 10.1109/MS.2013.85.

[35] A. E. Hassan and T. Xie, "Software intelligence: The future of mining software engineering data," in *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research*, ser. FoSER '10, New York, NY, USA: ACM, 2010, pp. 161–166, ISBN: 978-1-4503-0427-6. DOI: 10.1145/1882362.1882397.

[36] C. Hibbs, S. Jewett, and M. Sullivan, *The Art of Lean Software Development: A Practical and Incremental Approach*, 1st. O'Reilly Media, Inc., 2009, ISBN: 0596517319, 9780596517311.

[37] D. Hubbard, *How to Measure Anything: Finding the Value of Intangibles in Business*, 2nd. Wiley, 2010, ISBN: 9780470625699.

[38] E. A. Ichu, *The Role of Quality Assurance in Software Development Projects: Software Project Failures and Business Performance*. Germany: LAP Lambert Academic Publishing, 2012, ISBN: 3659169609, 9783659169601.

[39] "Ieee standard glossary of software engineering terminology," *IEEE Std 610.12-1990*, pp. 1–84, Dec. 1990. DOI: 10.1109/IEEESTD.1990.101064.

[40] A. Iqbal, O. Ureche, M. Hausenblas, and G. Tummarello, "Ld2sd: Linked data driven software development," in *In 21st International Conference on Software Engineering and Knowledge Engineering (SEKE 09*, 2009.

[41] I. Jacobson and E. Seidewitz, "A new software engineering," *Communications of the ACM*, vol. 57, no. 12, pp. 49–54, Nov. 2014, ISSN: 0001-0782. DOI: 10.1145/2677034.

[42] A. Jain and S. Angadi, "Gamifiying software development process," *Infosys Labs Briefings*, vol. 11, no. 3, pp. 21–28, 2013, http://goo.gl/T9PB96 accessed 01-Jan-2015.

[43] R. W. Jensen, "Improving software development productivity: Effective leadership and quantitative methods in software management," in, 1st. Upper Saddle River, NJ, USA: Prentice Hall Press, 2014, ch. 15. Function Point Sizing, ISBN: 0133562670, 9780133562675.

[44] C. Jones, *Applied Software Measurement: Assuring Productivity and Quality*, 2nd. Hightstown, NJ, USA: McGraw-Hill, Inc., 1997, ISBN: 0-07-032826-9.

[45] C. Jones, *Software Engineering Best Practices*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 2010, ISBN: 007162161X, 9780071621618.

[46] C. Jones. (Jun. 2012). Scoring and evaluating software methods, practices, and results. Namecook Analytics Blog, [Online]. Available: http://goo.gl/3i06pN.

[47] C. Jones, "The technical and social history of software engineering," in, 1st. Addison-Wesley Professional, 2013, ch. 10, ISBN: 0321903420, 9780321903426.

[48] C. Jones. (Jul. 2013). Why "cost per defect" is harmful for software quality. Namecook Analytics Blog, [Online]. Available: http://goo.gl/QUsvlw.

[49] M. Jørgensen, "A strong focus on low price when selecting software providers increases the likelihood of failure in software outsourcing projects," in *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*, ser. EASE '13, New York, NY, USA: ACM, 2013, pp. 220–227, ISBN: 978-1-4503-1848-8. DOI: 10.1145/2460999.2461033.

[50] M. Jorgensen, "What we do and don't know about software development effort estimation," *IEEE Software*, vol. 31, no. 2, pp. 37–40, Mar. 2014, ISSN: 0740-7459. DOI: 10.1109/MS.2014.49.

[51] C. Kaner and W. P. Bond, "Software engineering metrics: What do they measure and how do we know?" In *METRICS 2004*, IEEE CS Press, 2004.

[52] R. S. Kaplan and D. P. Norton, "The balanced scorecard: Measures that drive performance," *Harvard Business Review*, pp. 71–80, Jan. 1992.

[53] R. S. Kaplan and D. P. Norton, "Using the balanced scorecard as a strategic management system," *Harvard Business Review*, Jul. 2007. [Online]. Available: http://goo.gl/4v871V.

[54] M. Klubeck, *Metrics: How to Improve Key Business Results*, 1st. Berkely, CA, USA: Apress, 2011, ISBN: 1430237260, 9781430237266.

[55] M. Kutner, C. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, 4th, ser. The McGraw-Hill/Irwin Series Operations and Decision Sciences. McGraw-Hill Higher Education, 2003, ISBN: 9780072955675.

[56] T. O. A. Lehtinen, M. V. Mäntylä, J. Vanhanen, J. Itkonen, and C. Lassenius, "Perceived causes of software project failures - an analysis of their relationships," *Journal Information and Software Technology*, vol. 56, no. 6, pp. 623–643, Jun. 2014, ISSN: 0950-5849. DOI: 10.1016/j.infsof.2014.01.015.

[57] E. Letier and C. Fitzgerald, "Measure what counts: An evaluation pattern for software data analysis," in *2013 1st International Workshop on Data Analysis Patterns in Software Engineering (DAPSE)*, IEEE, 2013, pp. 20–22.

[58] S. Maheshwari and D. C. Jain, "A comparative analysis of different types of models in software development life cycle," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, no. 5, pp. 285–290, May 2012, ISSN: 2277-128X.

[59] A. Marcus and T. Menzies, "Software is data too," in *Proceedings of the FSE/SDP Workshop on Future of Software Engineering Research*, ser. FoSER '10, New York, NY, USA: ACM, 2010, pp. 229–232, ISBN: 978-1-4503-0427-6. DOI: 10.1145/1882362.1882410.

[60] T. McCabe, "A complexity measure," *IEEE Transactions on Software Engineering*, vol. SE-2, no. 4, pp. 308–320, Dec. 1976, ISSN: 0098-5589. DOI: 10.1109/TSE.1976.233837.

[61] T. Menzies, B. Caglayan, Z. He, E. Kocaguneli, J. Krall, F. Peters, and B. Turhan. (Jun. 2012). The promise repository of empirical software engineering data, [Online]. Available: http://promisedata.googlecode.com.

[62] T. Menzies and T. Zimmermann, "Goldfish bowl panel: Software development analytics," in *2012 34th International Conference on Software Engineering (ICSE)*, Jun. 2012, pp. 1032–1033.

[63] J. P. Miguel, D. Mauricio, and G. Rodríguez, "A review of software quality models for the evaluation of software products," *International Journal of Software Engineering & Applications (IJSEA)*, vol. 5, no. 6, pp. 31–54, Nov. 2014, http://www.airccse.org/journal/ijsea/papers/5614ijsea03.pdf.

[64] A. B. M. Moniruzzaman and S. A. Hossain, "Comparative study on agile software development methodologies," *Global Journal of Computer Science and Technology*, vol. 13, no. 7, 2013. [Online]. Available: http://arxiv.org/abs/1307.3356.

[65] P. Naur and B. Randell, Eds., *Software Engineering: Report of a Conference Sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968, Brussels, Scientific Affairs Division, NATO*. 1969, http://homepages.cs.ncl.ac.uk/brian.randell/NATO/nato1968.PDF.

[66] J. Olson, *Data Quality: The Accuracy Dimension*, ser. The Morgan Kaufmann Series in Data Management Systems. Elsevier Science, 2003, ISBN: 9780080503691.

[67] D. Parmenter, *Key Performance Indicators: Developing, implementing, and using winning KPIs*. Hoboken, New Jersey: John Wiley and Sons, Inc., 2010.

[68] L. H. Putnam and W. Myers, *Five Core Metrics: The Intelligence Behind Successful Software Management*. New York, New York: Addison-Wesley Professional, 2013.

[69] R. Ramler and K. Wolfmaier, "Economic perspectives in test automation: Balancing automated and manual testing with opportunity cost," in *Proceedings of the 2006 International Workshop on Automation of Software Test*, ser. AST '06, New York, NY, USA: ACM, 2006, pp. 85–91, ISBN: 1-59593-408-1. DOI: 10.1145/1138929.1138946.

[70] J. Raynus, *Software Process Improvement with CMM*. Norwood, MA, USA: Artech House, Inc., 1999, ISBN: 0-89006-644-2. [Online]. Available: http://goo.gl/Jc7Krf.

[71] J. Rost and R. Glass, "The dark side of software engineering: Evil on computing projects," in. Wiley, 2011, ch. 2 Lying, ISBN: 9780470922873.

[72] W. W. Royce, "Managing the development of large software systems: Concepts and techniques," in *Proceedings of the 9th International Conference on Software Engineering*, ser. ICSE '87, Los Alamitos, CA, USA: IEEE Computer Society Press, 1987, pp. 328–338, ISBN: 0-89791-216-0.

[73] V. Rubin, C. W. Günther, W. van der Aalst, E. Kindler, B. F. Van Dongen, and W. Schäfer, "Process mining framework for software processes," in *Proceedings of the 2007 International Conference on Software Process*, ser. ICSP'07, Berlin, Heidelberg: Springer-Verlag, 2007, pp. 169–181, ISBN: 978-3-540-72425-4.

[74] G. Ruhe and F. Gesellschaft, "Knowledge discovery from software engineering data: Rough set analysis and its interaction with goal-oriented measurement," English, in *Principles of Data Mining and Knowledge Discovery*, ser. Lecture Notes in Computer Science, J. Komorowski and J. Zytkow, Eds., vol. 1263, Springer Berlin Heidelberg, 1997, pp. 167–177, ISBN: 978-3-540-63223-8. DOI: 10.1007/3-540-63223-9_116.

[75] N. B. Ruparelia, "Software development lifecycle models," *SIGSOFT Softw. Eng. Notes*, vol. 35, no. 3, pp. 8–13, May 2010, ISSN: 0163-5948. DOI: 10.1145/1764810.1764814.

[76] A. Saltelli, S. Tarantola, and F. Campolongo, "Sensitivity analysis as an ingredient of modeling," *Statistical Science*, vol. 15, no. 4, pp. 377–395, 2000. [Online]. Available: http://projecteuclid.org/euclid.ss/1009213004.

[77] G. Snijkers, G. Haraldsen, J. Jones, and D. Willimack, *Designing and Conducting Business Surveys*, 2nd ed., ser. Wiley Series in Survey Methodology. John Wiley & Sons, 2013, ISBN: 9781118447918.

[78] W. B. Snipes, "Evaluating developer responses to gamification of software development practices.," Master's thesis, North Carolina State University, 2013. [Online]. Available: http://repository.lib.ncsu.edu/ir/handle/1840.16/9199.

[79] I. Sommerville, *Software Engineering*, 6th ed. Harlow, England: Addison-Wesley, 2001.

[80] S. L. Spraragen, "The challenges in creating tools for improving the software development lifecycle," in *Proceedings of the 2005 Workshop on Human and Social Factors of Software Engineering*, ser. HSSE '05, New York, NY, USA: ACM, 2005, pp. 1–3, ISBN: 1-59593-120-1. DOI: 10.1145/1082983.1083118.

[81] R. Swanstrom. (Mar. 2015). Cri scores for dissertation, Sense, [Online]. Available: http://goo.gl/NuEZsf.

[82] R. Swanstrom. (Mar. 2015). Dissertation-scoring-sdo, Github, [Online]. Available: https://github.com/ryanswanstrom/dissertation-scoring-sdo.

[83] Q. Taylor and C. Giraud-Carrier, "Applications of data mining in software engineering," *International Journal of Data Analysis Techniques and Strategies*, vol. 2, no. 3, pp. 243–257, Jul. 2010, ISSN: 1755-8050.

[84] D. Tosi, L. Lavazza, S. Morasca, and D. Taibi, "On the definition of dynamic software measures," in *Proceedings of the ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, ser. ESEM '12, New York, NY, USA: ACM, 2012, pp. 39–48, ISBN: 978-1-4503-1056-7. DOI: 10.1145/2372251.2372259.

[85] F. F. Tsui, *Essentials of software engineering*, 3rd. Jones & Bartlett Publishers, 2013.

[86] W. van der Aalst, *Process Mining : Discovery, Conformance and Enhancement of Business Processes*. Heidelberg: Springer, 2011.

[87] W. van der Aalst, "Process mining: Overview and opportunities," *ACM Transactions on Management Information Systems*, vol. 3, no. 2, 7:1–7:17, Jul. 2012, ISSN: 2158-656X. DOI: 10.1145/2229156.2229157.

[88] M. van Genuchten, R. Mans, H. Reijers, and D. Wismeijer, "Is your upgrade worth it? process mining can tell," *IEEE Software*, vol. 31, no. 5, pp. 94–100, Sep. 2014, ISSN: 0740-7459. DOI: 10.1109/MS.2014.20.

[89] K. Werbach, "(re)defining gamification: A process approach," English, in *Persuasive Technology*, ser. Lecture Notes in Computer Science, A. Spagnolli, L. Chittaro, and L. Gamberini, Eds., vol. 8462, Springer International Publishing, 2014, pp. 266–272, ISBN: 978-3-319-07126-8. DOI: 10.1007/978-3-319-07127-5_23.

[90] V. Winter, C. Reinke, and J. Guerrero, "Sextant: A tool to specify and visualize software metrics for java source-code," in *Emerging Trends in Software Metrics (WETSoM), 2013 4th International Workshop on*, May 2013, pp. 49–55. DOI: 10.1109/WETSoM.2013.6619336.

[91] T. Xie, S. Thummalapenta, D. Lo, and C. Liu, "Data mining for software engineering," *Computer*, vol. 42, no. 8, pp. 55–62, 2009, ISSN: 0018-9162. DOI: 10.1109/MC.2009.256.

[92] D. Zhang, Y. Dang, J.-G. Lou, S. Han, H. Zhang, and T. Xie, "Software analytics as a learning case in practice: Approaches and experiences," in *Proceedings of the International Workshop on Machine Learning Technologies in Software Engineering*, ser. MALETS '11, New York, NY, USA: ACM, 2011, pp. 55–58, ISBN: 978-1-4503-1022-2. DOI: 10.1145/2070821.2070829.

[93] D. Zhang, S. Han, Y. Dang, J.-G. Lou, H. Zhang, and T. Xie, "Software analytics in practice," *IEEE Software*, vol. 30, no. 5, pp. 30–37, Sep. 2013, ISSN: 0740-7459. DOI: 10.1109/MS.2013.94.