Abstract

Title of Dissertation: An Approach to Improving Existing Measurement Frameworks in Software Development Organizations

Manoel Gomes de Mendonça Neto, Doctor of Philosophy, 1997

Dissertation directed by: Professor Victor R. Basili
Department of Computer Science

Measurement is a key mechanism to characterize, evaluate, and improve software development, management, and maintenance processes. Nowadays, software organizations use metrics for very different purposes. Data is collected to describe, monitor, understand, assess, compare, validate, and appraise very diverse attributes related to software processes or products. Improving data collection and better using the existing data are important problems for software organizations.

This dissertation proposes an approach for improving measurement and data use when a large number of diverse metrics are already being collected by a software organization. The approach combines two methods. One looks at an organization's measurement framework in a top-down fashion and the other looks at it in a bottom-up fashion.

The top-down method, based on the Goal-Question-Metric (GQM) Paradigm, is used to identify the measurement goals of data users and map them to the metrics being used by the organization. This allows the measurement practitioners to: (1) identify which metrics are and are not useful to the organization; and (2) check if the goals of data user groups can be satisfied by the data that is being collected by the organization.

The bottom-up method is based on a data mining technique called Attribute Focusing (AF). It is used to identify useful information in the existing data that the data users were not aware of.

To validate the approach and to assess its usefulness, a case study was performed in a real industrial environment. The top-down and bottom-up methods were applied in the customer satisfaction measurement framework at the IBM Toronto Laboratory. The top-down method was applied to improve the customer satisfaction (CUSTSAT) measurement from the point of view of three data user groups. The bottom-up method was used to gain new insights into the existing CUSTSAT data.

The top-down method identified several new metrics for the interviewed user groups. It also contributed to better understanding the data user needs and led to modification of some of the data analyses and presentations done for those groups. The bottom-up method produced important insights on both the customer satisfaction domain and the measurement framework itself. Unexpected associations between key variables prompted new insights on their importance for the organization. Some of these associations have also revealed problems with the metrics being used to collect the data.

An Approach to Improving Existing Measurement Frameworks in Software Development Organizations

by

Manoel Gomes de Mendonça Neto

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
1997

Advisory Committee:

Professor Victor R. Basili, Chairman/Advisor
Professor Claude E. Walston
Professor Ben Shneiderman
Professor Marvin V. Zelkowitz
Dr. Inderpal S. Bhandari

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Measurement is a key mechanism to characterize, evaluate, and improve software development, management, and maintenance processes. Nowadays, software organizations use metrics for very different purposes. Data is collected to describe, monitor, understand, assess, compare, validate, and appraise very diverse attributes related to software processes or products.

Much of the research on software engineering measurement has dealt with the definition and validation of software engineering metrics and models [10, 38, 58, 51, 89]. Several works have also dealt with the problems of planning and implementing measurement programs in software organizations [61], most notably on goal-oriented measurement [17, 40, 88]. However, very little attention has been given to the problem of improving existing measurement programs.

This dissertation proposes an approach for improving measurement and data use when a large number of diverse metrics are already being collected by a software organization. The approach combines two methods. One looks at an organization's measurement framework in a top-down fashion and the other looks at it in a bottom-up fashion. These methods are used to: (1) better understand the data user needs, (2) evaluate how well the data that are being collected can fulfill those needs, and (3) extract new and useful information from the already existing data.

The approach was experimentally validated in a case study run in a real industrial environment. The proposed approach, the case study results, and lessons learned in improving an existing measurement program are the main contributions of this dissertation.

1.1 Problem Statement and Work Motivation

Metrics are not used in isolation. We define a **measurement framework** (MF) as a set of related metrics, data collection mechanisms, and data usage inside a software organization.

Software organizations measure for a reason. They usually have specific information needs. They assemble measurement frameworks to fulfill those needs. In general, software organizations have evolved their measurement frameworks over time, based upon input from a variety of sources and needs, but without a well structured set of goals. This scenario can lead to poorly structured measurement and data use. Software organizations lose their global understanding of the data (and its usefulness) in large and poorly structured measurement frameworks.

It is not uncommon to find software organizations that are: (1) collecting insufficient data; (2) collecting redundant data; (3) collecting data that nobody uses; or (4) collecting data that might be useful to people that do not even know it exists inside their organization. This makes measurement more expensive and, what is worse, less effective in fulfilling the data users needs. For these reasons, improving on-going measurement is an important problem for many software organizations.

1.2 Overview of Solution

We believe that the solution for this problem needs to address three key issues: (1) better understand the on-going measurement; (2) better structure it; and (3) better explore the data that the organization has already collected. Our work proposes an approach that addresses these three critical issues jointly.

1.2.1 Work Objectives

Our work does not intend to be a comprehensive or definitive approach to improve measurement frameworks. The objective of our work is to provide and validate an integrated set of techniques for:

O1- discovering interesting data distributions and associations in the MF database

O2- visualizing data distributions and associations in the MF database

O3- assessing the importance of metrics for specific user groups and for the organization as a whole

O4- assessing the structure (i.e., measurement instrument, scale, and domain value) of metrics used in the MF

O5- assessing the appropriateness of the data collection process

O6- assessing the importance of data analyses for specific user groups and for the organization as a whole

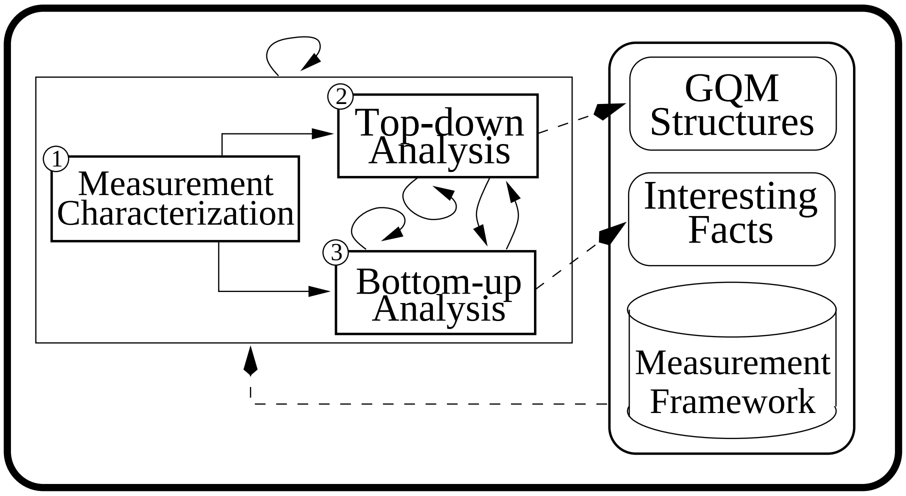

Figure 1.1: The Approach

O7- understanding and documenting the needs of users with respect to existing metrics, data analyses, and data presentations

O8- understanding and documenting the measurement goals of the MF data users

O9- identifying new applications and user groups for the data

O10- identifying the need for new metrics, data analyses, and data presentations

1.2.2 Overview of Approach

The approach combines a knowledge discovery technique, called Attribute Focusing (AF), with a measurement planning approach, called the Goal–Question–Metric Paradigm (GQM). In this approach, a characterization process (1) is used to understand on-going measurement. A GQM-based method (2) is used to structure it. And, an AF-based method (3) is used to discover new interesting information in the existing data. The approach is depicted in Figure 1.1.

The first phase – characterization – is executed to identify the (current and prospective) data user groups and how they are (or could be) using the data. The second phase – top-down analysis – is based on the GQM paradigm. It is executed to capture the goals of the data users and to map these goals to the metrics and data in the measurement framework (MF). In this way, one can detect what type of data is missing and what data is not being used in a MF. The third phase – bottom-up analysis – is based on the AF technique. It is executed to extract knowledge (useful, interesting, and non-trivial information) from the already existing data.

Figure 1.1 shows the control flow (using solid lines) and the information flow (using dashed lines) of this process. The dashed lines show the two main products of our approach: (1) GQM structures, produced by the top-down analyses; and (2) interesting facts, produced by the bottom-up analyses. The solid arrows indicate interactions between the phases. The characterization results are used to execute the bottom-up and top-down analyses. Thus, the characterization can be seen as a pre-requisite for the other two phases. The top-down and bottom-up phases can interact with each other. Interesting facts discovered during bottom-up analyses can lead to new measurement goals for the top-down analyses. Measurement goals can in turn be used to define new data sets for the bottom-up analyses.

1.3 Work Validation

We do not claim that our approach completely fulfills all the objectives listed in Section 1.2.1. Our work validation aims to:

1. Evaluate the degree to which our approach fulfills those objectives.
2. And, determine if those objectives are really important for improving a measurement framework.

Ultimately, we want to answer the following validation question:

1. Do the benefits of applying our approach compensate for its cost ?

In order to answer this question, we decided to validate our work experimentally through a case study. The approach was applied in an industrial environment and the results were analyzed to evaluate the approach's cost effectiveness.

1.3.1 Why Experimental Validation ?

Computer science is a relatively new field that has evolved mostly from mathematical sciences. It has inherited a strong tradition of analytical research from this discipline. Software engineering – as a branch of computer science – has inherited this bias towards analytical research. However, software engineering methods and tools are especially difficult to study analytically:

- One usually does not have well founded theories associated with software engineering technologies – notable exceptions are formal methods and design of programming languages.

- There are no universal laws or theories to model human factors associated with the people that apply those software engineering technologies.

These difficulties are also present in our research. There are very few models and theories associated with software engineering measurement, and there is no work on (much less theories associated with) the use of methods to improve existing measurement frameworks. For these reasons, experimental was chosen over analytical validation.

1.3.2 Why a Case Study ?

There are several experimental methodologies to validate new software technologies [113]. In the case of our approach, one might consider executing a small replicated experiment in an artificial setting, a controlled large scale experiment, or a case study in an industrial setting. In order to validate the assumption that the approach was useful to improve large measurement frameworks, it was decided that the approach should be applied in a real industrial environment. This decision discarded the use of a small replicated experiment in an artificial setting. The use of a controlled large scale experiment was discarded because it was impractical. Several industrial scale measurement frameworks in similar settings would be needed to do that. The chosen validation method was to execute a case study [74] in which the approach was to be applied to a real industrial measurement framework and its results compared with the existing ad-hoc process to improve this measurement framework.

1.4 Experimental Platform

We applied our approach to improve the Customer Satisfaction (CUSTSAT) Measurement Framework at the IBM Toronto Laboratory. The CUSTSAT data is collected annually by surveys carried out by an independent party. Its purpose is to evaluate customer satisfaction with products of IBM's Software Solutions Division and their competitors. The IBM Toronto Laboratory is only one of the several IBM Software Solutions laboratories that use the CUSTSAT data. Inside the IBM Toronto Laboratory, the CUSTSAT data is used by several different groups (e.g., development, service, support, and senior management).

IBM surveys a large number of customers from several different countries. All the data is stored in one database. Currently, this database already stores several years of CUSTSAT data. The large amount of data, the diversity of groups that are interested in it, and the maturity of this measurement framework made it a very good platform to validate our approach for improving measurement frameworks.

1.5 Main Contributions

This work gives some important contributions to the software engineering and data mining fields:

- In software engineering, the contributions are:
  - The design of the case study used to evaluate the methods.
  - The process defined to characterize existing measurement frameworks.
  - The instantiation of the GQM Paradigm to improve existing measurement frameworks.
  - The formalization of some important GQM concepts, such as:
    - the semantic of the facets of a measurement goal.
    - the templates for GQM questions.
- In data mining, the contributions are:
  - The association of data mining methods with the AF Technique, in particular:
    - the use of generic relationship questions to create attribute classes to reduce the space searched by the AF Technique.
    - the use of attribute ordering to improve visualization of cause-effect relations in the AF diagrams.
    - the use attribute classes to define an algorithm to organize the AF diagrams.
  - The use of generic relationship questions to create an interface between GQM (a measurement planning paradigm) and AF (a data mining technique).

1.6 Definitions

This section introduces the terminology and acronyms used throughout this dissertation. The terminology adopted here was adapted from the data mining terminology proposed by Klösgen and Zytkow [75] and the software engineering measurement terminology proposed by Fenton [52]. During this section (and the rest of this dissertation), **boldface font** is used when new terms are defined.

We define **application domain** as the real or abstract system a software organization wants to analyze using a MF. An **entity** (object, event, or unit) is a distinct member of an application domain. Similar entities can be grouped into classes such as persons, transactions, locations, events, products, and processes. Entities are characterized by attributes and relations to other entities. An **attribute** (field, variable, feature, property, magnitude) is a single characteristic of all entities in a particular entity class, for instance "usability" of software products or "size" of source code. In the case of a measurement framework, an attribute defines "what" one wants to measure. A **relation** is a set of entity tuples which has a specific meaning, for instance "a is married to b" (for person entities "a" and "b"). We measure entity attributes to empirically define relations between entities, for instance we can determine the relation "a is heavier than b" by weighing entities "a" and "b."

Measurement is the process of assigning a value to an attribute. A **metric** is the mapping model used to assign values to a specific attribute of an entity class. A metric states "how" we measure something. It usually includes a measurement instrument, a value domain, and a scale. **Data** is a set of measured (collected, polled, surveyed, sensed, observed) attribute values produced by specific metrics for certain user groups.

A **user group** is a formal group inside the organization that in some way utilizes (consumes, employs) the data produced by the MF. A **data use** is a description of the way a user group consumes the data. And, a **data user** is any member of a user group. A **data manager** is a person responsible for managing the collection and storage of, and/or access to the data in a measurement framework. A person may play both roles – data manager and data user – in a given MF.

A measurement **goal** is an operational, tractable description of a user group objective in using the data. In this dissertation, a goal is always described using the template we will introduce in Section 2.3.3. **Domain knowledge** is non-trivial and useful empirical information specific to the application domain believed to be true by the data users. **Background knowledge** is the domain knowledge that data users had before analyzing the data. And, **new** or **discovered knowledge** is the new domain knowledge that data users gain by analyzing the data.

The following acronyms will be used throughout this dissertation:

- MF: measurement framework.
- GQM: goal-question-metric.
- AF: attribute focusing.
- MC: measurement (framework) characterization.
- CUSTSAT: customer satisfaction.
- SQ: CUSTSAT survey question.

- DA/P: data analysis or data presentation.
- SWS: IBM Software Solutions Division.
- Toronto Lab: IBM Toronto Laboratory.

1.7 Outline

The rest of this dissertation is organized as follows. Chapter 2 describes background material related to our work. The use of goal-oriented measurement in software organizations is described. Data mining and machine learning from software engineering data is discussed. Special sections describe the Goal-Question-Metric (GQM) paradigm and the Attribute Focusing (AF) technique.

The approach is introduced in Chapter 3. The process that is executed during each of the approach's phases is described in detail. Section 3.1 describes the process used to document the key components of a measurement framework. Section 3.2 describes the GQM-based method used to capture and map data user goals to the metrics of a MF. Section 3.3 describes the AF-based method used to extract new knowledge from data available on a MF.

Chapter 4 describes how the approach was applied to the IBM's Customer Satisfaction Measurement Framework. The MF components documented during the characterization process, the GQM structures produced by the top-down analysis, and the interesting facts obtained from the bottom-up analyses are discussed there. Chapter 5 presents the approach validation. The criteria used to validate the approach, the validation results, and the approach evaluation are discussed there. Chapter 6 suggests future research opportunities and discusses the main results, contributions, and limitations.

Chapter 2

Background and Literature Review

This dissertation starts with the premise that a good measurement framework should be sound, complete, lean, and consistent. A MF is **sound** when its metrics and measurement models are valid in the environment where they are used. A MF is **complete** when it measures everything that its users need to achieve their goals. A MF is **lean** when it measures what is needed and nothing else (metrics cost money to collect [42]). A MF is **consistent** when its metrics are consistent with the user goals. This means that: (1) the metrics scale and range of values are suitable for the user needs; and (2) the metrics can be applied when and where they are needed by the users.

Requiring soundness, completeness, leaness, and consistency of measurement frameworks is not new a new idea in software measurement. In a seminal 1976 work, Boehm, et al. [112], wrote:

> "Our ... approach were as follow: 1. Determine a set of characteristics which are important ... and reasonably exhaustive and non-overlapping. ... 3. Investigate the characteristics and associated metrics to determine their correlation with software quality ... 4. Evaluate each candidate metric ... and ... its interactions with other metrics: overlaps, dependencies, shortcomings, etc."

Although, all four issues were identified early by measurement practitioners, most of the work published on measurement validation is concerned with the issue of using sound metrics.

Metrics have been validated in very different ways. Analytical validations have been done: (1) to analyze if a metric is theoretically sound [43, 50, 83, 114]; or (2) to verify if a metric fulfills the properties that are associated with the attribute it is supposed to measure [2, 31, 102, 108, 111]. Empirical validations of predictive models have been done to validate these models' precision and accuracy [27, 32, 71, 109]. Empirical validation of direct metrics has been done: (1) to analyze the association between these metrics and important quality measures [9, 16, 19, 72, 100]; and, (2) to assess these metrics consistency when they are used by different people to measure the same thing [73, 100].

There are few works on the validation of MFs' completeness, leanness, and consistency. These three issues have traditionally been addressed in practitioner's examples of successful MFs [40, 60, 91, 96, 110]. Only recently, methodologies have been proposed to build complete, lean, and consistent MFs [61, 88]. Most of these works recognize that measurement should be executed in a top-down goal-oriented way, but they only address the problem of defining lean, complete, and consistent MFs. Little attention has been given to the problem of improving the completeness, leanness, and consistency of existing operational MFs. This dissertation deals precisely with this issue.

2.1 Choosing an Approach for Quality Improvement

Measurement is not an end in itself. A software organization measures to establish a quantitative and qualitative basis to improve software quality and cost. In other words, measurement should be integrated in a larger framework that supports understanding, assessment, improvement, packaging, and reuse of experiences (knowledge, processes, technologies, and methods) in software organizations. To this end, this section examines some of the organizational approaches used to improve quality in various types of business.

2.1.1 Total Quality Management

The goal of Total Quality Control (TQM) is to generate institutional commitment to success through customer satisfaction — the term was coined to describe the Japanese management style to quality improvement [49]. The approaches to achieve TQM vary greatly in practice. In general, however, they seek to achieve total quality of a product by involving all members of the production process in the improvement effort.

TQM was developed in Japan based on the ideas of W. E. Deming [44] and J.M. Juran [69]. The principles of TQM were successfully applied in industries for mass production, such as automobile and consumer electronics industries. In those industries, the concept of total customer satisfaction was translated in terms of producing parts and products with zero defect. Statistical process control — a periodical random sample of products — was used to assess and control the quality of the production.

In software organizations, the concept of total quality is not so clear cut. It is difficult to define and evaluate the quality of software products. It is difficult capture software customer needs. It is practically impossible to remove all faults from a software product. Those difficulties are added to the fact that each software product is complex, abstract, and unique [33, 34]. The success histories of TQM in manufacturing industries could not be easily transferred to the software industries [3, 57], not even in Japan [68, 70].

2.1.2 The Lean Enterprise Management

The Lean Enterprise Management (LEM) goal is to build a product using the minimal set of activities and materials needed, eliminating non essential steps and costs. LEM has been used to improve factory output. Womack, et al. [112], have written a book on the application of LEM to automotive industries.

LEM basic idea is to tailor a process suited to the product needs. Given the characteristics for a product V, it selects the appropriated mix of sub-processes $p_1, \ldots, p_N$ to satisfy the goals for V, yielding a minimal tailored process P_V to produce V.

$$Process(P_V) \longrightarrow Product(V) \tag{2.1}$$

The ideas of LEM are very useful in software development as software organizations have to learn from one process about another, and the development process has to be tailored to each new product that is developed [14].

2.1.3 The Plan-Do-Check-Act Approach

The Plan-Do-Check-Act (PDCA) is a quality improvement process based upon a feedback cycle for optimizing single process production lines. Its based on work by W. A. Shewhart [103] and was made popular and applied effectively to improve Japanese manufacturing after World War II by W. E. Demming [44]. The approach is defined as four basic steps:

> Plan: Develop a plan for improving the existing production process. Set up quality targets (using measurable criteria) and methods to achieve the targets.
>
> Do: Carry out the plan complying with development standards and quality guidelines.
>
> Check: Observe the plan effects at each stage of development against the quality criteria set up at the planning phase.
>
> Act: Study results to determine what was learned, what problems occurred, and what can be improved in the next cycle.

The PDCA basic idea is to produce an improvement cycle over the $Process(P)$ used to produces a $Product(X)$. The PDCA cycle produces a family of processes $\{P_i\}$ and a series of product versions $\{X_i\}$. Each cycle introduces an modification in a process P_j in order to improve it over process P_{j-1}. The improvement cycle is experimentally checked by examining if X_j has better quality than X_{j-1}.

$$Proc(P_1), Proc(P_2), \ldots, Proc(P_N) \longrightarrow Prod(X_1), Prod(X_2), \ldots, Prod(X_N) \tag{2.2}$$

The PDCA idea of creating process improvement cycles has been adapted to software organizations [3, 45, 68]. However, the process improvement cycle is more complex in software industries. Each software product is unique and requires its own process. In software, each improvement cycle has to build a "new" process tailored from previous software development experiences [18].

2.1.4 The SEI Capability Maturity Model

The Capability Maturity Model (CMM) [63, 90] is a quality improvement approach that was specifically tailored to Software Development. CMM is based on the idea of quality management maturity models developed by Likert [77] and Crosby [39]. The idea of using a software maturity model was developed by Radice while at IBM [95] and was made popular by Humprey at the Software Engineering Institute (SEI) [66].

CMM uses a five-level process maturity model to improve quality (Table 2.1). A maturity level is defined based on repeated assessment of an organization's capability in key process areas. Improvement is achieved by action plans for processes that had a poor assessment result. The SEI has developed a process improvement cycle to support the movement through process levels. Basically, it consists of the following activities:

- Initialize:
  - Establish sponsorship
  - Create vision and strategy
  - Establish improvement structure
- For each maturity level:
  - Characterize current practice in terms of key process areas
  - Assess recommendations
  - Establish improvement strategy

CMM Level	Focus	Key Process Areas
1. Initial	Key people and heros	
2. Repeatable	Processes for project management	Requirements management, project planning, tracking, oversight, quality assurance, subcontract management, and configuration management
3. Defined	Engineering processes	Process definition, training program, integrated management, product engineering, intergroup coordination, peer reviews
4. Managed	Product and process quality	Quantitative process and quality management
5. Optimizing	Continuous process improvement	Defect prevention, management of technology, and process changes

Table 2.1: CMM Maturity Levels

- For each key process area:
  - Establish process action teams
  - Implement tactical plan, define processes, plan and execute pilot(s), plan and execute institutionalization
  - Document and analyze lessons
  - Revise organizational approach

This process does not examine the product or any other business characterization. It assumes that there are essential or idealized processes and that adhering to these processes will generate good products. In CMM, a $Process(P_I)$ that is in level "I" is modified on key areas until this process is at level "$I+1$." Hopefully, $Process(P_{I+1})$ will produce better products than $Process(P_I)$.

2.1.5 The Quality Improvement Paradigm

This dissertation adopts the 'Quality Improvement Paradigm' (QIP) as the basis for improvement of software quality and productivity. The **Quality Improvement Paradigm** (**QIP**) is a long-term, quality-oriented, meta-lifecycle model for software organizations [14, 18]. The QIP promotes understanding, assessing, and packaging of software development experiences as the means to improve software quality.

Figure 2.1 describes the QIP. It highlights the two learning cycles of the QIP Paradigm: the intra-project monitor-control cycle, and the inter-project (corporate) learn-improve cycle.

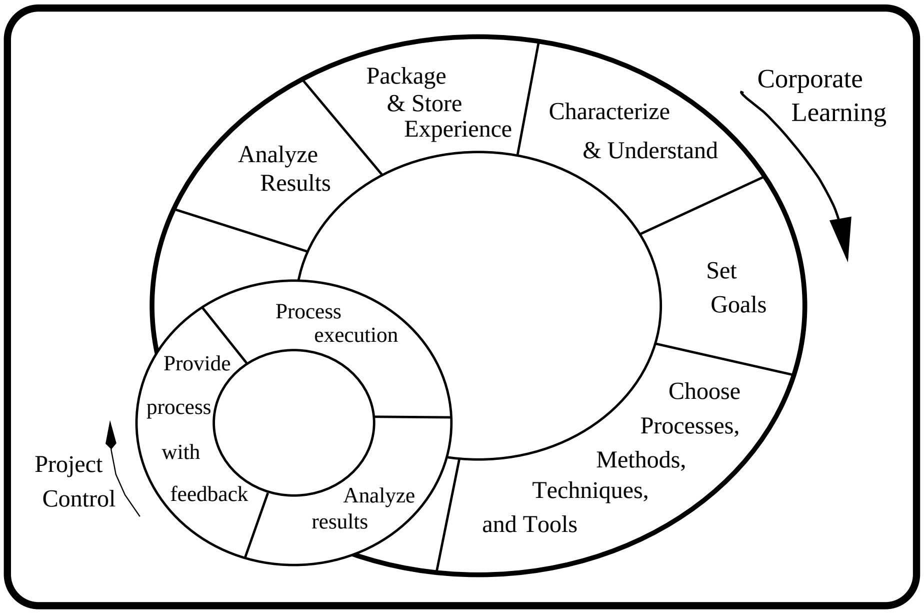

Figure 2.1: The Quality Improvement Paradigm

The QIP evolved from the lessons learned in the NASA Software Engineering Laboratory [11, 12, 15, 17]. In its current form the QIP has six essential phases:

1. **Characterize the environment** - this involves understanding a software project and its context qualitatively and quantitatively so that the correct decisions can be made.

2. **Set goals** for project and organization - this consists of a process of setting goals and decomposing them into detailed subgoals. The process works interactively until it has produced subgoals that we can measure directly.

3. **Choose and tailor a process model** that satisfies the goals - this involves selecting and tailoring the life cycle, methods, techniques and tools to satisfy the project goals relative to the characterized environment.

4. **Execute the process** - this involves the construction of products according to the process model chosen in the previous phase. The data prescribed by phases 2 and 3 is collected validated and used to keep the process under control. The collected data allow us to monitor the process and take contingency actions when necessary (feedback).

5. **Analyze the collected data** during and after the project - this phase is partially done during the process execution for project control as described in phase 4. It is also done post-mortem to better understand the nature of

software development and check what experience can be gained from each project.

6. **Learn and feedback** during and after the project - intra-project feedback is done during project execution to solve project contingencies as described in phase 4. Inter-project feedback is based on phase 5. It is done post-mortem by packaging experiences into models and other forms of structured knowledge that can be reused in the future.

The QIP incorporates ideas from several quality improvement approaches used in the manufacturing industries [14]:

- Its evolutionary nature, based on feedback loops, is similar to the Plan-Do-Check-Act Paradigm (PDCA) [44, 103].
- Its goals, feedback mechanisms, and use of measurement allow us to involve everyone in the job of quality assurance. One can use QIP to implement a Total Quality Management (TQM) philosophy in a software organization [49].
- Its approach to tailoring the development process as an optimum set of available sub-processes is similar to Lean Enterprise Management (LEM) [112]. Both have the idea of meeting the particular goals of a project using the minimum set of essential steps.

What basically differentiates the QIP from the LEM, TQM and PDCA approaches is that the QIP is tailored to software development, while the others were essentially used to improve the quality of an assembly-line-like production environment [14]. The difference is that each software product is unique. In software development, one has to learn from one process about another, the quantitative models are less rigorous and more abstract, and the development process has to be tailored to each new product that is developed [7].

QIP basic idea is to tailor a process suited to the project needs based of the goals stated for this project. Given the project goals and quality requirements for a product V, it selects the appropriated mix of sub-processes $p_1, \ldots, p_N$ to satisfy the goals for V, yielding a tailored process P_V to produce V.

$$Process(P_V) \longrightarrow Product(V) \tag{2.3}$$

The sub-process $p_1, \ldots, p_N$ used to build $Process(P_V)$ are drawn from the organization experiences. They are built upon understanding the relationships between the historical projects and products and the goals for the new $Product(V)$.

In terms of being tailored to quality improvement of software organizations, only the CMM [90] can be compared to the QIP. The Experimental Software Engineering Group at Maryland has adopted the QIP because we believe that an organization should focus on the specific problems they want to solve [82]. Unlike the CMM, the QIP does not assume that process improvement is dependent on the maturity of the organization [106]. The QIP starts with a CMM level 5 style of organization, even though it does not have level 5 capability yet [6]. The organization is driven by the understanding of its business, products and process problems [14]. It learns from its own business, not from an external generic process model.

2.1.6 Mapping our Improvement Approach to QIP

The QIP can been seen as a framework for applying the scientific method to software organizations. Our experimental work can be mapped to the QIP. The characterization of IBM's CUSTSAT MF corresponds to the QIP first phase. From IBM's point of view, the case study has the goal of improving its measurement framework. From this dissertation point of view, the case study has the goal of evaluating the improvement methods. The MF top-down and bottom-up analyses correspond to the approach execution cycle. The validation of our approach presented in Chapter 5 corresponds the result analysis phase. The improved MF and a packaged set of improvement processes is the final result of using this paradigm.

2.2 Looking at Measurement Frameworks in a Top-down Fashion

Measuring for software quality and cost improvement is not a simple task [59]. The cost and quality of software products are associated with its development process as opposed as to its production[1] process [93]. Software is an abstract and complex product and software development is a human intensive process[34, 33].

One of the key ideas behind our approach is that, in software organizations, measurement should be defined in top-down goal-oriented fashion. Gilb put it better when he said [55]: "Projects without clear goals will not achieve their goals clearly." A variety of goal-oriented measurement paradigms have appeared in the literature: the Quality Function Deployment [76] (QFD) ; the Software Quality Metrics [81] (SQM); and the Goal Question Metric (GQM) paradigm are some of them.

[1]Software production corresponds to the act of recording a software product and its installation procedures into the storage media to be shipped to a customer.

Required Quality				Quality Attributes				
Primary	Secondary	Tertiary	Importance	A1	A2	A3	A4	A5
Req. 1	Req. 1.1	Req. 1.1.1	1	x		x		
		Req. 1.1.2	1	x				
	Req. 1.2	Req. 1.2.1	3	x	x			
		⋮				⋮		
Req. N	Req. N.1	Req. N.1.1	2				x	
		Req. N.1.2	2					x
		Req. N.1.3	3					x

Table 2.2: QFD Matrix

2.2.1 Quality Function Deployment

The Quality Function Deployment (QFD) [1, 76] is a technique that evolved from the TQM principle of deriving measures from the customer point of view. In QFD, quality requirements are established for each product from the customer point of view. Those requirements are then mapped to metrics to be used to satisfy those customer needs. This mapping is done using "House of Quality" matrices. Table 2.2 shows an abstract example of such matrix (adapted from [76]). The left side of the matrix has the captured customer requirements. Those requirements are refined top-down – "deployed" – in more detailed ones. The detailed requirements are ranked by importance (in this cases from 1 to 3) from the customer point of view. The right side of the matrix map each of those requirements to the measurable attributes that will be used to evaluate them.

2.2.2 Software Quality Metrics

Software Quality Metrics (SQM) was developed to allow the customer to assess the product being developed by a contractor [28, 81]. In SQM, a set of quality factors is defined for the final product. Those factors are refined into a set of criteria (attributes), which are mapped to a set of pre-defined metrics. In this way, the SQM user just selects the factors and criteria of interest to define which metrics will be used to assess the delivered product. Figure 2.2 shows an example of a SQM structure adapted from [28].

2.2.3 The Goal-Question-Metric Paradigm

The Goal-Question-Metric Paradigm was proposed by Basili [20, 17] as a means of measuring software in a purposeful way. The GQM paradigm first step is to

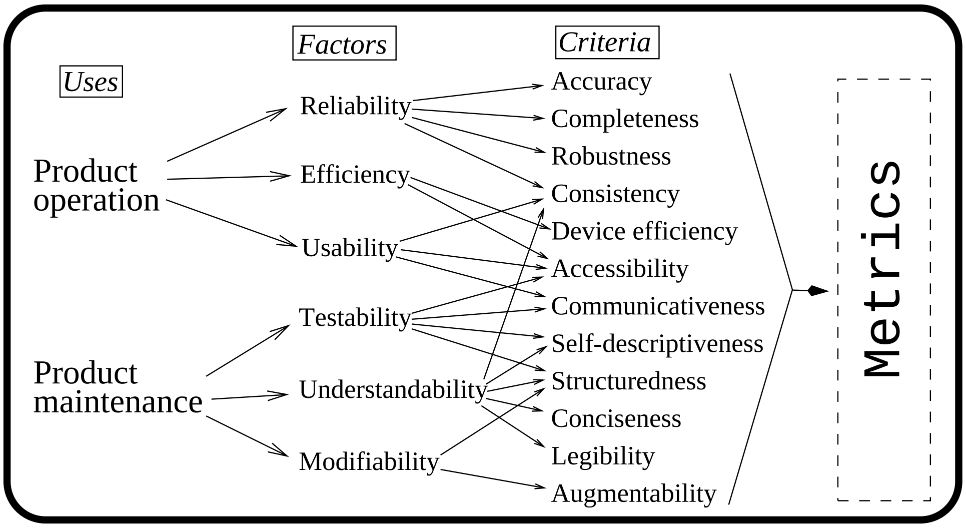

Figure 2.2: A SQM structure

define measurement goals tailored to the specific needs of an organization. Goals are refined in a operational, tractable way, into a set of quantifiable questions. Questions in turn imply a specific set of metrics and data for collection. This paradigm has been used successfully in several organizations (e.g., NASA [15], Motorola [40], HP [58], AT&T [4]).

Figure 2.3 shows an abstract example of what we call a GQM structure. The following template — defined by Basili and Rombach [17] — is used to define measurement goals:

> Analyze <u>'object of study'</u> in order to <u>'purpose'</u> with respect to <u>'focus'</u> from the point of view of <u>'point of view'</u>. (2.4)

Each of the underlined words above represents a facet, that must be considered in measurement planning. For example:

> Analyze 'service support for our product' in order to 'evaluate it' with respect to 'customer satisfaction' from the point of view of 'service support personnel'. (2.5)

Each goal implies several questions based on its facets. For example, the purpose "evaluate" might generate questions of the type: "How does the service support of our product compare with its competitors ?" or "How does the current service support satisfaction compare with previous years ?"

The questions will then be refined into the metrics needed. The goal facets are also used in this process. For example, the point of view determines the scale, granularity and timing of the metrics used to answer a certain question.

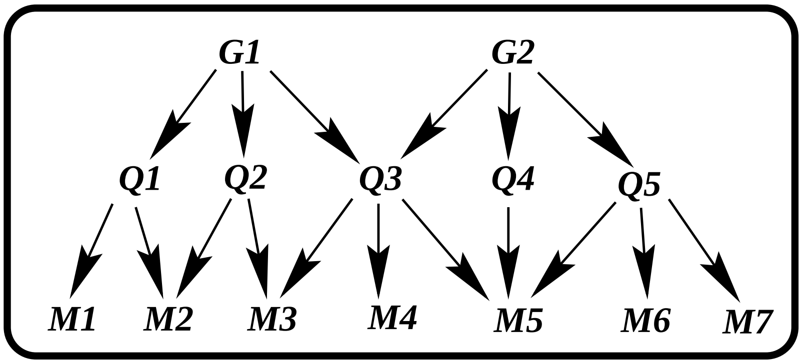

Figure 2.3: An abstract GQM structure

GQM has a wider scope than QFD and SQM. The QFD has been used for the purpose of planning and control the final product. The SQM has been used for the purpose of assessing the final product. Both were conceived to measure things from the point of view of customers and users. To the GQM Paradigm, the measurement point of view, object of study, and purpose are input variables. In fact, SQM and QFD can be considered subsets of the GQM Paradigm [13]. This dissertation adopts the Goal-Question-Metric Paradigm [20, 17] to improve MFs in a top-down fashion.

2.3 Instantiating the GQM Paradigm

The GQM is a general paradigm that has been instantiated in several different ways [4, 8, 15, 46, 58]. All those instantiations aim to define measurement from scratch. This dissertation will use its own instantiation of the GQM Paradigm. Instead of being tailored to define new MFs from scratch, our "version" is tailored to improve existing MFs.

Figure 2.4 shows the type of GQM structure that will be built in this dissertation. This type of structure provides a platform to interpret the data and better understand the data user needs. It specifies the goals associated with a certain data user group (goals with the same "point of view"). Those goals will be refined in attributes (i.e., what the data users want to measure), and those attributes will be mapped to the metrics that are being used in the MF. In this way, the structures will allow data managers to trace the goals of a certain data user group to the measures that are intended to define them operationally. Looking at Figure 2.4 for example, one may conclude that metrics $M2$, $M3$, and $M5$ are useful to the data user group. Metrics $M1$ and $M4$ are extraneous to them. An the metrics to measure attributes $A3$ and $A5$ are missing altogether from the MF.

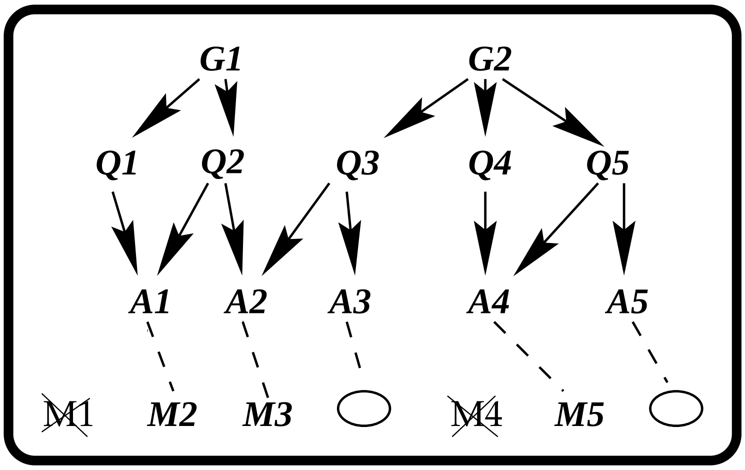

Figure 2.4: GQM Structure for the Top-down Analyses

The process to build GQM structures will be described in Chapter 3. The rest of this section describes the way we will describe metrics, attributes, questions, and goals in this dissertation.

2.3.1 Expressing Metrics

In Section 1.6, a metric was defined as a mapping model used to assign values to an attribute of an entity class. This definition recognizes the fact that a measure is a mapping from the empirical (real) world to a formal (mathematical) world [53, 62, 73, 97]. A metric is characterized by a measurement model, a value domain, and a scale. In order to describe a metric, one has to identify all those components.

The metric **value domain** is the set of values that can possibly be assigned by a metric to the attribute it is measuring. These values are represented by suitable symbols (usually numbers). The definition of a value domain consists in declaring the set of symbols used to represent the measured values.

The value domain is not enough to define what operations can be executed over these values. A measurement **scale** is needed to define what operations are admissible over a value domain. The work on scales was pioneered by Stevens [105], the following four scale types make up his original classification:

- **Nominal scale** defines representations that can be used to classify entities into categories based on their attribute values. Examples of attributes that can be expressed in a nominal scale are: sex of a person, race of a person, type of a software fault.

Scale Type	Description	Example
Nominal Scale	establishes categories	sex, race, type of a software fault
Ordinal Scale	establishes rank orderings	rank in class, level experience of a programmer
Interval Scale	establishes the notion of unit	temperature in ^oC or ^oF, calendar time of a project
Ratio Scale	establishes unit and absolute zero	temperature in ^oK, elapsed time of a project

Table 2.3: Types of Measurement Scale

- **Ordinal scale** augments the nominal scale by placing a logical ordering in the attribute classifications. Examples of attributes that can be expressed in a ordinal scale are: rank in class, level of experience of a programmer, seriousness of a software fault.
- **Interval scale** augments the ordinal scale by introducing the notion of unit into the classifications, i.e. the differences between levels of attributes values on any part of the scale reflect equal differences in the attributes measured. Examples of metrics that are expressed in an interval scale are: temperature in degrees Celsius and degrees Fahrenheit, and calendar time of a project.
- **Ratio scale** augments the interval scale by introducing the notion of absolute zero, i.e. the absence of the attribute measured. Examples of metrics that are expressed in a ratio scale are: temperature in degrees Kelvin, number of lines of code, and elapsed time of a project.

Although there can be other scale types, the categories listed on Table 2.3 are the most common and cover almost all software engineering measures. Those four categories will be used in this dissertation to identify the measurement scales.

The metric **measurement model** is the procedure, instrument, or function used to associate a value to an entity attribute. For example, the size of C programs may be measured by the metric "lines of code." In this case, the measurement model is the exact procedure used to count the "lines of code" – e.g., count the number of non-comment semi-colons in the "source code." This metric value domain is the natural numbers in ratio scale.

In the above example, "lines of code" is said to be a **direct metric** because its values are directly derived from an entity. In the other end of the spectrum, there are the **indirect metrics**. These metrics uses a function to determine an attribute value from other attribute values. They are commonly used in software engineering resource and prediction models ([10] part II and [38] chapter 6). For example, Boehm's Cost Constructive Model [27] (a well known effort prediction model) uses the following relation to measure the attribute "Effort:" $E = 2.4 \times \left(\frac{LOC}{1000}\right)^{1.05}$ In this case, the measurement model is the "E" function and the value domain is the real numbers in ratio scale.

2.3.2 Expressing Attributes

As defined in Section 1.6, attributes state what one wants to measure. Examples of attributes used in software organizations are: complexity, size, coupling, and cohesion of software code, type and seriousness of a defect, experience and capability of programmers. Although the differentiation between "metrics" and "attributes" was made popular by Fenton in his 1991 software metrics book [51], it was first proposed by Rubey and Hartwick in a seminal 1968 work [98], on their own words:

> "... attribute is a precise statement of a specific software characteristic. ... a metric was developed for quantitative measurement of each quality attribute. These metrics ... can be used to produce a numerical value that makes it possible to compare a given program with other programs or a desired standard."

Well defined attributes allow people to check if the MF has sound metrics to measure these attributes. This dissertation recognizes basically three types of attribute description: the implicit, textual, and formal descriptions. **Implicit descriptions** occur when an attribute is described only by its name. For example, suppose that a organization wants to measure the attribute "project staff size." This attribute may be (implicitly) described by its name if this is enough to establish the attribute meaning. In the example, common sense might be enough to determine that the "number of programmers in a project" measures the "staff size."

Often, however, the description of what one wants to measure is not clear by the attribute name alone. In the previous example, it might not be clear what one means by "staff" (e.g., are secretaries part of the project staff ? what about the acceptance testers ?). In such cases, a **textual description** might be used to define what one wants to measure. For example, "the project staff size should include all the people that have worked at some point in designing, coding, reusing, testing, and maintaining software for the project."

Sometimes even the textual description of what an attribute tries to measure can be confusing, especially for software product attributes such as size, complexity, and coupling. In such cases, a **formal description** should be used to describe this attribute. This dissertation endorses the use of property-based approaches [31] to formally define an attribute. The property-based attribute definition works by stating as axioms the properties that the metrics used to measure the attribute have to satisfy.

Consider the following axiom stated by Weyuker as a property of the attribute "code complexity" [111]: The concatenation of a program body "R" with two different programs bodies "P" and "Q" can affect complexity in different ways. Although "R" has a fixed complexity in isolation, "R" may interact with "P" in subtly different ways than it interacts with "Q." It can produce different levels of complexity when concatenated with "P" and "Q." This can be formally stated as: $(\exists P)(\exists Q)((\mathcal{M}(P) = \mathcal{M}(Q)) \wedge (\mathcal{M}(R;P) \neq \mathcal{M}(R;Q)))$, where $\mathcal{M}(P)$ means measured complexity of "P," and "$R;P$" means concatenation of program "R" and "P." In her paper [111], Weyuker showed that the cyclomatic number [80] — a very common code complexity metric — does not satisfy this property.

The property-based approach can be used to describe attributes, because it isolates the attribute definition from the metric definition. One can discuss whether the attribute "complexity" should have a property, independent of whether some metric really satisfies this property or not. Once one has agreed on the properties of an attribute, one can validate the metrics used to measure it. For example, if one agrees that the property in previous example is a property of the "code complexity," then the "cyclomatic number" cannot be validated as a complexity measure. Note that the attribute properties are by no means a complete "description" of the attribute. They only state a formal basis upon which to validate the metrics used to measure these attributes.

2.3.3 Expressing Goals

In general terms, a goal can be defined as a planned position or result to be achieved. This dissertation will call them general goals to differentiate them from our very specific definition of a 'goal'. From the business management point of view, one can classify general goals into two large domains [5]: 'strategic goals' and 'organizational goals'.

- Strategic goals are general goals affecting the nature the business in which a firm engages (e.g. a database software company can have as a strategic goal to enter the workstation DBMS market).

- Organizational goals are general goals affecting the way that parts of the corporation are organized and the production process executed (e.g. a database software company can have as a organizational goal to produce more robust DBMS systems).

In software development organizations, strategic goals are more related to the business management field, while organizational goals are more related to the software engineering management field. As expected, this dissertation focus on the latter. It uses organizational goals to guide the process of defining or improving a software measurement framework.

In software engineering measurement, 'goals' should state what is to be analyzed, from what perspective, and for which purpose. This dissertation uses the template introduced in Section 2.2.3 to state these goals:

$$\text{Analyze } \underline{\text{`object of study'}} \text{ in order to } \underline{\text{`purpose'}} \text{ with respect to } \underline{\text{`focus'}} \text{ from the point of view of } \underline{\text{`point of view'}}. \tag{2.4}$$

This dissertation calls an organizational goal set using the previous goal template a 'measurement goal', or simply a **goal**. Goals are defined in terms of purpose and perspective:

- The purpose outlines the object of study and what one wants to do with it.
- The perspective outlines what aspects of the object of study are relevant, and who is interested in such aspects.

Each one of the templates underlined words represents one facet. Facets are keywords that will substitute for the underlined words to produce a goal. The template's four facets (object of study, focus, purpose, and point of view) have very specific semantics.

Object of study is any entity in the organization of the template user.

Purpose is represented by one of the following keywords: characterize, assess, evaluate, control, improve, or predict.

Focus is the primary attribute one is interested in measuring.

Point of view is any data user group in the software organization.

The semantics of the facets "object of study" and "point of view" are explained by the definition of entity and data user group given in Section 1.6. Likewise, the semantics of the facet "focus" is explained by the definition of attribute given in the same section. The semantics of the "goal purpose" facet needs further explanation. This dissertation uses the following key words to express the goal purpose:

Characterize - define and select metrics to measure the attribute associated with the goal focus and measure them from the goal point of view.

Assess - use a predefined set of metrics to compare the object of study attributes against some predefined standard.

Evaluate - define and select metrics as before, derive baselines for these metric values (usually from experience, e.g. historical values), measure the attributes, and compare the obtained values against the baseline. Evaluate if the current values are better, similar, or worse than expected.

Control - define and select metrics as above, derive baselines for these metric values, measure the attributes, and compare values against the baseline. If attribute values are worse than the baseline, take corrective action to keep them within the prescribed bounds.

Improve - define and select metrics as above, derive baselines for these metric values. State improvement targets based on baseline values and take affirmative action to achieve these targets (e.g., improve the process, adopt a new technology, train people, etc.). Measure the attributes, and compare their values against the improvement targets. If necessary, take corrective actions to achieve the planned targets.

Predict - define and select a predictive measurement model and execute it. This indirect metric must be executable early in the software life cycle so that the data users can estimate up front a value for the attribute they are trying to measure (estimate). Follow up on the predictions and revise the estimates during the project execution.

2.3.4 Expressing Questions

Questions are used in the GQM Paradigm as the link between measurement goals and metrics. This dissertation splits questions in two orthogonal categories:

- 'Characterization questions'
- 'Relationship questions'

Characterization questions are used to define the attributes (and metrics) that will be measured to pursue the stated measurement goals. The characterization questions are influenced by the object of study and focus of the measurement goals. They aim to characterize the entities related to the objects of study. The following template is used to express characterization questions:

$$\text{What is the } \underline{\text{attribute X}} \text{ of } \underline{\text{entity Y}} \text{ ?} \tag{2.6}$$

For example:

What is the 'number of open requirements' in 'the project design phase' ? (2.7)

Relationship questions will define how the collected data will be analyzed to pursue the measurement goals. These questions are directly related to a goal purpose and point of view. They aim to state the relations between attributes of the object of study and attributes of other entities that one wants to investigate empirically. The following template is used to express relationship questions:

How does <u>attribute X</u> of <u>entity Y</u> relate to (affect, compare to) <u>attribute Z</u> of the <u>object of study</u> ? (2.8)

For example:

How do the '# of open requirements' in 'the project design phase' affect 'error proneness' of 'the final product' ? (2.9)

In this dissertation, the characterization questions will be used in the top-down analyses and the relationship questions will be used in the bottom-up analyses. Chracterization questions are used to define attributes in the GQM-based method described in Section 3.2.1. Relationship questions are used in the first step of the AF-based method described in Section 3.3.1.

Its is important to highlight that the use of questions templates to support the generation of questions in a GQM model is a new contribution of this dissertation to the GQM Paradigm.

2.4 Looking at a Measurement Frameworks in a Bottom-up Fashion

Existing or legacy data is the most important asset of any measurement framework. For this reason, improving data usage is one of the best ways to improve a MF as a whole. One of the key assumptions of this dissertation is that MF databases contain useful information that is not being explored by the data users. The bottom-up analysis aims explore a MF database to infer new useful information (knowledge) about the application domain and the MF itself.

2.4.1 Machine Learning

Information can be inferred from a database by deduction or induction [65]. Deduction infers information that is a logical consequence of the information in the database. This information is always true provided that the database contents

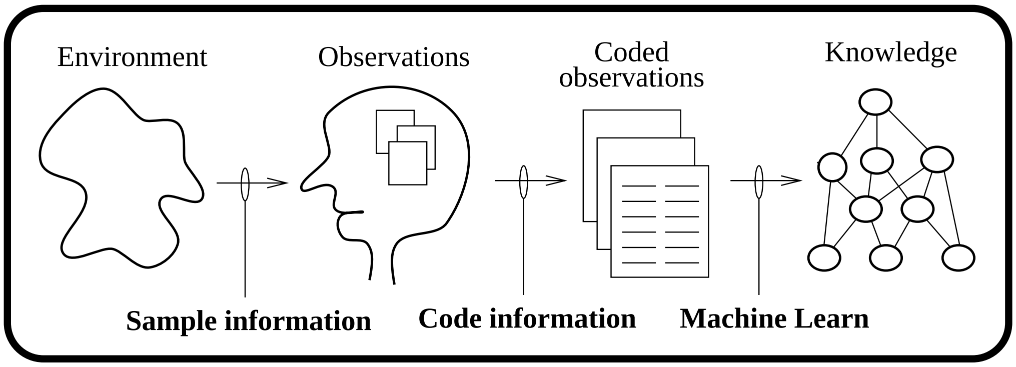

Figure 2.5: A Machine Learning Framework

are true. Induction infers information by generalization of the information contained in the database. This information is believed to be true and is supported by data patterns in the database. Consider a database with ten similar software systems developed and maintained by the same group of people. Suppose that they have measures of the development and maintenance costs of these systems, and each project used one of the following two languages: FORTRAN or Ada. In this case, one can deduce the average cost of the projects done in FORTRAN and in Ada. However, one can only induce information on which language is more costly to use in general.

The induced information will express general statements or rules about the entities being measured. It is usually higher-level information, that is not necessarily true, but it is believed to be true due to the contents of the database. If this information is also interesting and previously unknown, it called **discovered knowledge** or **new knowledge** [54]. The process of discovering new knowledge is referred to as **inductive learning** [64].

The automation of inductive learning processes has been researched in an artificial intelligence area called **machine learning** [99]. A machine learning system does not interact directly with its environment. It uses "coded observations" of this environment to learn about it. Figure 2.5 depicts the machine leaning process. It samples facts from the environment we want to model. It codes these facts as "coded observations" of the environment. These coded observations are fed into a machine learning mechanism to produce a model of the environment. The model can then be used to derive unknown and interesting information (i.e. new knowledge) about the environment [36, 65].

A broad range of machine learning approaches can be fit into the above framework. Candidate elimination algorithms [84], decision tree algorithms [87, 94], explanation-based algorithms [35, 41, 85], neural network algorithms [79, 86], and genetic algorithms[64], although very different, all fit this very general framework.

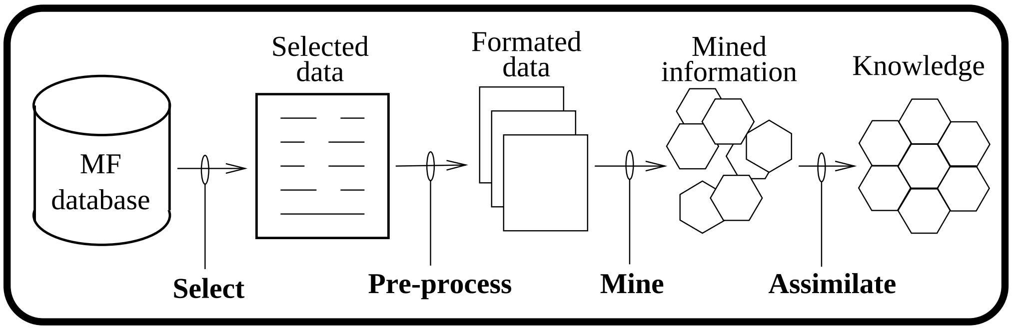

Figure 2.6: A Data Mining Framework

Let's consider two extreme examples:

- Neural networks use training sets which are coded observations of the environment. The environment model is represented in a neural network as patterns of interactions between the simple computational elements. The learning algorithm works by adjusting the weights and thresholds of the network connections. Knowledge is implicitly stored in the network itself as a vast number of connections and weights [79].

- Decision trees also use a set of training instances. The learning algorithm builds a classification tree as the environment model. The tree classifies examples among a finite number of classes. Nodes of the tree are labeled with attribute names, the edges are labeled with possible values for these attributes, and the leaves are labeled with the different classes [94].

Some machine learning techniques – such as decision tress – represent knowledge in an interpretable symbolic format. Other techniques – such as neural networks – represent knowledge implicitly in a non-interpretable format [99]. Most of the reported uses of machine learning techniques in software engineering use techniques that represents knowledge in symbolic interpretable format. Techniques such as Classification Trees [92, 101, 104, 107] and Optimized Set Reduction [29, 32] have been used more frequently than neural networks to build predictive and classification models for software organizations.

2.4.2 Data Mining

The bottom-up analysis aims to extract knowledge directly from the MF database. The research area that studies machine learning systems that draw "coded observations" directly from a database is called 'data mining' [48]. Formally, **data mining** is defined as the process of inducing previously unknown, and potentially useful, information from databases [54]. Figure 2.6 shows a data mining framework (adapted from [47]).

Although the framework for data mining and machine learning may seem similar, there is an important distinction. The database is designed for purposes others than data mining. The representation of entities and attributes in the database has been chosen to meet the needs of the applications that use it rather than the needs of data mining [65]. In our case, the database is designed to meet the needs of the measurement framework as stated by the software organization's goals. This means that the data is not organized in a way that will facilitate machine learning. In particular, there might be irrelevant, missing, noisy, and uncertain data in the database.

There are two basic types of data mining operations (see [67] for a more detailed classification): (1) one can data mine to create predictive and classification models to forecast the future (predictive data mining); or (2) one can data mine to discover interesting facts in a database (forensic data mining). The goal of the bottom-up analysis described in this dissertation is not to use the MF data to build models, but rather to extract new interesting facts from it. For this reason, the bottom-up analysis uses a forensic data mining technique.

Forensic data mining techniques are not completely automated. They usually involve one or more people during the assimilation phase of the data mining process (see Figure 2.6). These people are usually experts in the application domain and their role is to transform the mined information into knowledge. Typical forensic data mining techniques search databases for deviations from expected data patterns, unknown associations between variables, or interesting sequencing of attribute values.

2.4.3 The Attribute Focusing Technique

The forensic data mining technique used as the basis for the bottom-up analysis is called Attribute Focusing. Attribute Focusing (AF) has been used in several different applications – including software process measurement [24, 23, 26], customer satisfaction [25], and sports [22] data analyses.

The AF technique searches an attribute-value (measurement) database for interesting facts. An **interesting fact** is characterized by the deviation of attribute values from some expected distribution or by an unexpected correlation between values of a set of attributes. The facts are presented in easily interpretable bar chart diagrams. The diagrams are sorted by **interestingness level** — a numeric value calculated to quantify how interesting each diagram might be to an expert. The ordered diagrams are presented to the experts. Knowledge discovery takes place when the experts address the questions raised by the diagrams.

Figure 2.7 shows an example of an Attribute Focusing diagram. It was obtained from a real data set pertaining to a particular class of software prod-

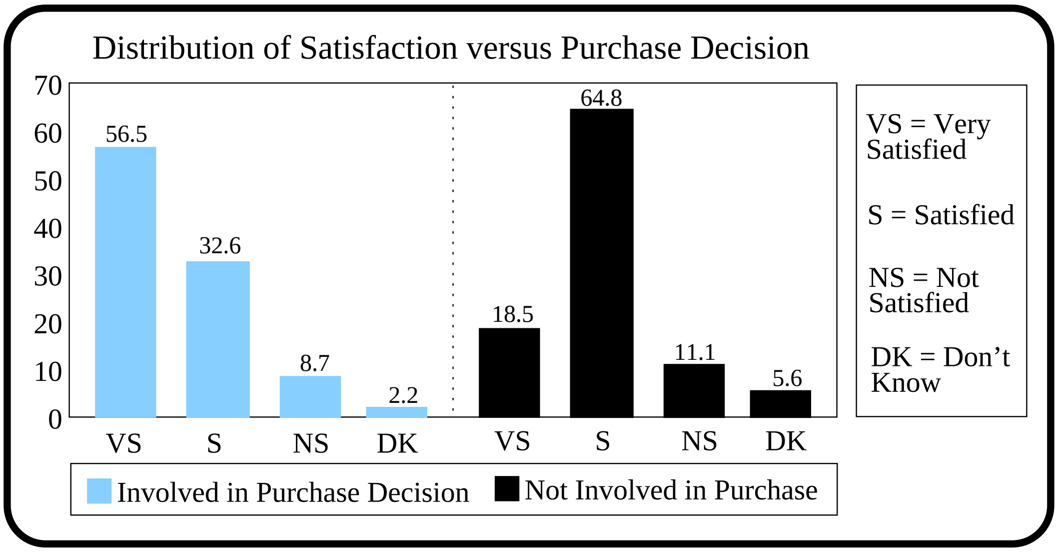

Figure 2.7: A Two-way Attribute Focusing Diagram

ucts [25]. Let us call it "Product Class X." This particular diagram has two attributes: "Overall Satisfaction" and "Customer Involvement in the Decision to Purchase the Product."

The satisfaction level by customer involvement in purchase is shown by bar patterns in the diagram. The possible values are: "involved in purchase decision," if the customer was involved in the decision to purchase the product he/she is evaluating, and "not involved in purchase," if not. The y-axis shows the percentage of occurrence of each "satisfaction" value per "purchase involvement" value. For example, the first vertical bar indicates that about 56.5% of those "involved in the decision to buy the product" were "very satisfied with the product."

The diagram in Figure 2.7 is saying that if the customer was involved in purchasing a product of Product Class X, he/she is likely to evaluate the product more favorably than customers that were not involved in the decision to buy this product (see the differences in values between "very satisfied" and "satisfied" for "involved" and "not involved in purchase decision").

This diagram exemplifies very well how the AF Tool helps knowledge discovery. It points out new facts to the experts. These facts may lead to discovered knowledge or not. The experts are the ones that will look at the facts expressed in the diagrams using their background knowledge, and conclude if the diagrams are saying something new and useful.

Suppose for example that the experts know that products of Class X are expensive (background knowledge). This might lead to the discovery that purchasers of this class of products try to defend the product in order to justify their decision to invest in it.

At this point, it is important to remark that the AF tool used in this dissertation was designed to mine nominal and ordinal data. The tool has limitations in exploring interval and ratio data. Numeric-valued attributes have to be mapped into discrete ranges of values before they can be used in an AF-analysis. Consider the metric "lines of code" (LOC) for example. Its numeric values have to be mapped to a discrete and finite set of values if one wants to use them in AF analyses. Suppose that the values "small, medium, and large" are considered adequate to quantify software size. In this case, one could define the metric "discrete software size" using the following mapping: (1) the discrete software size is "small" if $LOC < 5000$; (2) it is "medium" if $5000 \leq LOC < 100,000$; and (3) it is "large" if $LOC \geq 100,000$.

"Interestingness" Functions

The diagram presented in Figure 2.7 is said to be a 2-way diagram because it involves two attributes. The function used to calculate the interestingness level of a 2-way diagram – involving two attributes "A_x" and "A_y" in nominal or ordinal scale – is:

$$Interestingness(A_x, A_y) = \forall u \forall v \{\max [In_2(A_x = v; A_y = u)]\} \tag{2.10}$$

The "In_2" function quantifies the association of two particular values "v" of "A_x" and "u" of "A_y." It calculates the probability of co-occurrence of these values as if the attributes were independent ($Observed(A_x = v) \times Observed(A_y = u)$), and subtracts from it the rate of occurrence of the combination observed in the data ($Observed(A_x = v \wedge A_y = u)$):

$$\begin{aligned} In_2(A_x = v; A_y = u) \quad = \quad & |Observed(A_x = v) \times Observed(A_y = u) - \\ & Observed(A_x = v \wedge A_y = u)| \end{aligned} \tag{2.11}$$

$Observed(A_x = v)$ is the observed rate of occurrence of value v over all A_x values, and $Observed(A_x = v \wedge A_y = u)$ is the rate of occurrence of value pair (v, u) over all (A_x, A_y) values.

Other interestingness functions can be used with the AF-technique. We have used functions that estimate the interestingness level for associations between an arbitrary number of attributes (N-way analysis).

2.4.4 Generic Relationship Questions

In the AF technique, it is very important to avoid the computation of uninteresting relations whenever possible. An uninteresting relation wastes machine time to compute, and yields uninteresting diagrams that will waste data user's time during the diagram reviews.

The AF Tool avoids uninteresting metric combinations based on user-defined data (metric) groups. The metrics grouped together are not correlated during the analysis. For example, a typical 2-way AF analysis will use two metric groups. The AF Tool will pick one metric from each group and try to correlate them in pairs.

This dissertation introduces the idea of using a 'generic relationship question' to select and group the data for the AF analyses. A relationship question — as described in Section 2.3.4 — can be used to state what relations between attributes one wants to investigate empirically. The AF-based method introduced in Chapter 3 will investigate several empirical relations in each analysis. It will use a **generic relationship question** (GRQ) to state the set of relations to be investigated empirically. The following template is used to define a GRQ:

$$\text{How do } \underline{\text{`Attribute class } X_1\text{'}} \text{ and ... and } \underline{\text{`Attribute class } X_{N-1}\text{'}} \\ \text{[relate to, affect, impact] } \underline{\text{`Attribute class Y'}} \text{ ?} \tag{2.12}$$

Attribute classes are sets of attributes grouped according to certain criteria or features relevant to a data user group. For example, attributes that represent logical features of the final products could be grouped in class "Y," while attributes representing managerial constraints over the project could be grouped in class "X_1." In this example, the above template would result in the following question for AF analysis:

How do the 'managerial constraints over the project' relate to
the 'logical features of the final products' ? (2.13)

Multiple attribute classes are used in the GRQ template (2.12) in order to make it suitable to define N-way analysis.

The generic relationship questions are a generalization of the relationship questions introduced in Section 2.3.4. GRQs effectively establish a link between the GQM Paradigm and the AF technique.

2.4.5 A Final Word on Interestingness

Let us go back to the concept of "interestingness." In a two-way AF analysis, an attribute association (i.e., a correlation between attribute values) is selected for presentation if:

$$Interestingness(A_x, A_y) > C, \\ \text{where C is a fixed interestingness cutoff value} \tag{2.14}$$

The above formalism of a two-way relationship can be extended into a three-way relationship. Let $In_3(A_x = v; A_y = u; A_z = t)$ be:

$$\begin{aligned} In_3(A_x = v; A_y = u; A_z = t) &= |Obs(A_y = u \wedge A_z = t) \times Obs(A_x = v) - \\ &\quad Obs(A_x = v \wedge A_y = u \wedge A_z = t)| \end{aligned} \tag{2.15}$$

A three-way relationship is interesting (with A_x as the focus attribute) if the absolute value of the association is greater than any of the 2-way relationships and also greater than the cutoff C. In other words:

$$\begin{aligned} Interestingness(A_x, A_y, A_z) &> Interestingness(A_x, A_y) \wedge \\ Interestingness(A_x, A_y, A_z) &> Interestingness(A_x, A_z) \wedge \\ Interestingness(A_x, A_y, A_z) &> C \end{aligned} \tag{2.16}$$

The cutoff value C is set by the expert or the person analyzing the data. It expresses the threshold for which an attribute association is considered interesting. Implicit to this algorithm is the selection of the strongest associations between attributes of an existing set of attributes. Also, implicit to this algorithm is the selection of the optimum length of the associations. That is, the three-way evaluation compares the results of the various two-way evaluations, if the three-way evaluation is greater than (more interesting than) the two-way evaluations, this pattern is output, otherwise not. This implies that the AF algorithm for N-way associations will converge to an optimally sized association description. Because the description of the pattern includes those and only those attributes associated with the focus attribute, the expert need not evaluate or compare other associations that differ by the presence or absence of other attributes since these comparisons have been automatically evaluated by AF. The convergence to optimal length associations is therefore essentially a convergence towards the most informative and parsimonious associations. A deeper discussion of these concepts can be found in [22, 37].

Although the strength and optimal length of an association is a statistical measure of interestingness, it may not be a complete measure of practical utility. A strong association is useful only if it is unexpected or previously unknown.

In the AF-based method that will be introduced in Section 3.3.1, generic relationship questions (GRQs) are used to declare which (unexpected or unknown) attribute associations should be considered in AF analyses. The attribute classes defined by the GRQs are used to incorporate the expert's domain knowledge to the AF search for interesting and useful associations.

The last and possibly most important step in the search for interesting patterns is to convey the discovered "interesting" information to the experts. In the AF-based method, diagrams are used to convey information to the experts. The attribute classes defined by the GRQs are used to organize the AF diagrams before they are shown to the experts (see algorithm in Page 49). The organization of the diagrams helps to create a more meaningful context in which the "interesting" attribute associations will be interpreted. Here, like before, the GRQs are used to increase the potential interestingness of the mined information.

In this dissertation, the experts that will revise the AF diagrams are the several data users and managers of the Customer Satisfaction Measurement Framework. They are assumed to be "lay people" on data analysis, but experts in their knowledge domain. For them, the interesting diagrams are those that involve unexpected or unknown associations between satisfaction attributes of their interest. The space to be searched for these associations is specified by generic relationship questions involving attribute groups of their interest. One of the major contributions of this dissertation is exactly to use the GRQs to incorporate the expert's domain knowledge to the AF search for interesting attribute associations.

Chapter 3

The Improvement Approach

The approach is composed of three phases. The first phase – characterization – is executed to identify the data user groups and how they are using the data. The second phase – top-down analysis – is based on the GQM paradigm. It is executed to build GQM structures for each data user group and use these structures to detect missing and extraneous metrics in a MF. The third phase – bottom-up analysis – is based on the AF technique. It is executed to extract knowledge from the data that already exists in the MF.

3.1 Measurement Framework Characterization

The measurement framework characterization (MC) is executed to identify the "key components" of a MF and document how they relate to each other. The key components we want to identify are: the metrics, attributes, existing data, user groups, and data uses.

The approach uses a combination of structured interviews and review of the available MF documents to capture and document those key components. A structured interview is one in which the questions are in the hands of the interviewer and the response rests with the interviewee (as opposed to an unstructured interview in which the interviewer simply raises topics for discussion and the interviewee provides both the relevant questions and the answers) [78]. Structured interviews are used to capture the descriptions of user groups, attributes, and data uses, as those MF components are usually not documented and can only be obtained by interviewing data managers and data users.

Reviews of the MF documents and database are used to obtain the descriptions of metrics and available data. Metrics are explicitly or implicitly described in measurement manuals and data collection forms, or any organization document that has the description of the data that should be collected through the MF. The available data can usually be identified by reviewing the descriptions of the MF database, or, in the worst case, by directly examining its contents.

3.1.1 The Characterization Process

We use the following process to characterize a measurement framework:

Step 1 — Identify Metrics

The first components to be identified are the metrics used in a MF. All the metrics used to collect data in a MF must be listed, including how they work – especially their measurement instrument, scale and value domain.

- Entry Criteria: none.
- Input: Available measurement documents.
- Procedure: List all metrics. For each metric, answer the following questions:
  - What is the measurement instrument ?
  - What is the metric scale and range of values ?
- Methodology: Review available measurement (or software process) documents
- Output: Description of metrics.
- Exit Criteria: All metrics are described and continued investigation reveals no additional metrics.

Step 2 — Identify Available Data

The second type of component to be identified is the data available in the MF. This includes when and under what circumstances the metrics were used to collect data, where the resulting data was stored, and how to access it. This last step may require an understanding of the format in which the data is stored and how to get authorization to use it. This may require a sizable amount of work if the data is stored in several different formats and/or locations.

- Entry criteria: Metrics description is available.
- Input: Metrics description and data repositories (e.g., MF database) documentation.
- Procedure: For each data repository, answer the following questions:
  - When and under what circumstances were the data collected ?
  - How is the collected data stored ?

  - How can the data be accessed ?
  - What data is available in the repository ?

- Methodology: Use metric descriptions and review data repositories documentation. Examine data repositories directly and interview the data managers responsible for the repositories if necessary.

- Output: Description of available data.

- Exit Criteria: All data repositories are described and continued investigation reveals no additional data repositories.

Step 3 — Identify Data Uses and User Groups

The third type of component to be identified are data uses. Each type of data analysis and presentation that is generated with the data must be described. Each description should include the frequency and granularity with which the data is used. Together with the data uses, who is using the data – the user groups – must also be identified. A user group description should include the objectives of the group in using the data as well as how important the data is for them.

- Entry Criteria: The metric and data descriptions are available.

- Input: Metrics and data descriptions.

- Procedure: List all data usage. For each metric and/or data group, describe the data analysis and presentations done with it. Each data usage description should answer the following questions:

  - How is the data used ?
  - What is the frequency of this data usage ?
  - What is the granularity at which the data is used ?
  - Who is using the data ?

  Based on the list of who is using the data, describe all data user groups. Each user group description should answer the following questions:

  - Who is the main representative of the data user group ?
  - What is the group's purpose in using the data ?
  - What is the user group role in the organization ?
  - How important is the data for the user group ?

- Methodology: Interview data managers. Talk to specific user group representatives if necessary.
- Output: Description of user groups and data uses.
- Exit Criteria: All user groups and data uses are described and continued investigation reveals no additional user groups or data uses.

Step 4 — Identify Attributes

The last components we have to describe are the attributes. This information is obtained by asking the user groups to describe their perception of what is being measured by the metrics. For example, when a person says a program size is 15,000 lines of code, this person is saying that the metric "number of lines of code" is being used to measure the attribute "size of a program."

- Entry Criteria: Descriptions of metrics and user groups are available.
- Input: Available measurement documents, metric and user group descriptions.
- Procedure: For each metric, describe precisely what the metric is supposed to measure.
- Methodology: Interview data user representatives or, if possible, extract the description directly from the MF documentation. Use one of the three types of attribute description discussed in Section 2.3.2
- Output: Description of attributes.
- Exit Criteria: All identified attributes are described and continued investigation reveals no additional attributes.

3.2 Top-down Analysis

This phase is used to capture the data user goals and to map them to the data that is being collected. This helps data managers gain a better understanding of the data user needs, and to identify missing and extraneous metrics in the MF. The top-down analysis is based on an instantiation of the Goal-Question-Metric Paradigm (Section 2.2.3).

As shown in Figure 2.4, this GQM-based method is applied to build (or revise) a structure that maps the data user goals to the metrics (and data) used in the organization. This structure is used to identify missing or extraneous elements of a MF.

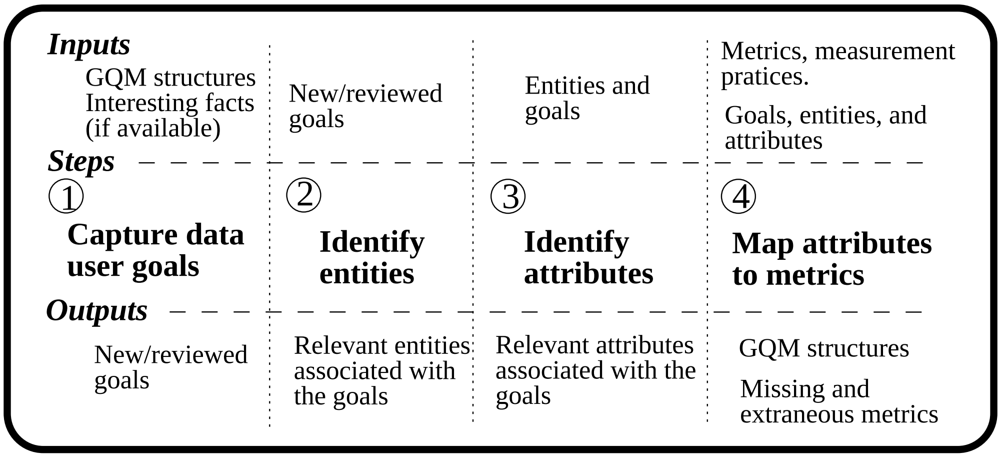

Figure 3.1: The GQM-based Method

3.2.1 The GQM-based Method

The GQM-based method applies the principles of the GQM paradigm to improve an existing MF. Our objective is to build a GQM structure in the mold of the one shown in Figure 2.4. Each structure is built by interviewing a representative of a data user group. This structure captures the measurement needs of this user group, and maps them to the existing metrics that are supposed to fulfill those needs. Figure 3.1 shows the process for building such GQM structures.

Step 1 — Capture Data User Goals

The first step of the method is to capture the goals of a user group. This is done by interviewing a data user group representative using the goal template described in Section 2.3.3. For each goal, the data user representative has to identify the measurement "object of study," "purpose," and "focus" (the "point of view" is the user group itself).

The "object of study" is the entity that the user group wants to analyze (e.g. product X). The focus is the primary attribute that the user group wants to measure in order to analyze that entity (e.g. customer satisfaction). The "purpose" outlines what the user group wants to do with the "object of study" (e.g. evaluate it). This dissertation recognizes the following measurement purposes: assess, characterize, evaluate, control, improve, and predict. The semantics of these words are explained in Section 2.3.3. A detailed textual explanation of the goal purpose should be captured from the interviewee.

This step is guided by the informal description of the data user objectives, obtained during the characterization process (see Section 3.1.1). Previous GQM structures and new knowledge about the data can also be used as input for this step, if they are available.

- Entry Criteria: Description of the group and its data uses are available.
- Input: Description of data user group objectives or previous GQM structures built for the group (if available), and new insights in the measured data (if available).
- Procedure: For each data use, ask the interviewee to list their goals in using the data in that way. For each goal answer the following questions:
  - What is the main object of study ?
  - What is the goal focus ?
  - What is the goal purpose ? (explain the goal purpose textually if necessary)
- Methodology: Interview the group representatives. Use the Goal Template 2.4 described in Section 2.3.3.
- Output: Updated list of the group goals in using the data.
- Exit Criteria: All measurement goals are described by the interviewee.

Step 2 — Identify Relevant Entities

The next step is to identify the entities whose attributes one wants to measure – called here **relevant entities**. The relevant entities can be identified in two ways: (1) asking about them during the interview with the representative of the data user group; or, (2) looking for them in the documentation available about the object of study.

Usually, two entities can directly be derived from each goal, one is the "object of study" itself and the other is the entity with which the "focus" attribute is associated. We identify other relevant entities by finding out which entities are related to the "object of study" and which may affect the "goal focus" from the data user group point of view.

Consider the Goal 2.5 listed in Section 2.2.3 as an example. There are two relevant entities listed in this goal: the service support process (object of study) and the customer (entity related with the goal focus). The other relevant entities might be: the product, the support team, the problem, the provided solution.

- Entry Criteria: Description of the group measurement goals is available.
- Input: measurement goals and documents associated with the objects of study (if available).

- Procedure: For each goal identify which entities are related to the object of study and goal focus. Create or (if possible) re-use existing lists of relevant entities for this user group.

- Methodology: Interview the group representatives and/or, if possible, extract the description directly from available documentation about the object of study.

- Output: Updated list of entities that are relevant to the group goals.

- Exit Criteria: All relevant entities are identified.

Until a detailed list of relevant entities is created for the user group, use the following abstract check list to identify relevant entities:

- If the examined entity is a product, consider as possibly relevant: (1) processes used to produce the product, (2) resources used to produce the product, (3) clients or users of the product, (4) models used to describe the product.

- If the examined entity is a process, consider as possibly relevant: (1) the personnel that enact the process; (2) products produced by the process; (3) products that are inputs for the process; (4) models used to describe the process.

- If the examined entity is a person or a group of people, consider as possibly relevant: (1) the organization unit in which the person or group of people is inserted; (2) the roles they play in their organization unit.

- If the examined entity is an organization unit, consider as possibly relevant: (1) the people that work in the organization; (2) the process the organization unit enacts; (3) the other organization units with which they interact; (4) their channels of interaction.

Step 3 — Identify Relevant Attributes

The next step is to identify the attributes one wants to measure to achieve this goal - called here **relevant attributes**. For each relevant entity, an initial list of entity attributes that might be relevant for the stated goal is prepared. The initial list of relevant attributes must be reviewed and expanded by the user group representative during an interview. The end result should be a list of attributes classified according to their relevance to the user group's goals.

- Entry criteria: description of the group measurement goals and relevant entities is available.

- Input: measurement goals, relevant entities, and documents associated with the objects of study (if available).
- Procedure: for each goal identify which attributes of the listed entities are relevant to the user group. Describe precisely each attribute and rate its importance to the user group goals.
- Methodology: if possible, extract the a comprehensive list of attributes directly from available documentation or use the existing attribute lists. Interview the group representatives to produce an updated list of relevant attributes. Ask them to rate the importance of each attribute. Describe the attribute in one of three ways discussed in Section 2.3.2 – implicitly, textually, or formally.
- Output: updated list of attributes that are relevant to the group goals.
- Exit criteria: all relevant attributes are identified.

In order to produce a comprehensive list of attributes for each entity, a checklist based on the entity type may be used, for example:

- If the entity is a **product**, consider:
  1. quality attributes of the product (number of defects, changes, stability, reliability, etc.)
  2. logical attributes of the product (functionality, capability, usability, maintainability, etc.)
  3. physical features of the product (size, complexity, modularity, coupling, etc.)
- If the entity is a **process**, consider:
  1. physical attributes of the process (size, complexity, etc.)
  2. managerial constraints over the process (budget, schedule, quality targets, etc.)
  3. process conformance (how well the process is performed)
- If the entity is a **person** or a **team**, consider:
  1. position/role in the organization
  2. professional motivation
  3. education level (training)

4. experience with the process and products they use (language, tools, virtual machine, etc.)
5. knowledge of the application domain

– If the examined entity is a **organization unit**, consider:

1. the primary business of the organization
2. physical attributes of the organization (size, complexity, gross expenditure, gross sales, number of software installations, type of platforms, number of employees, etc.)
3. management features of the organization (rate of innovation, expenditure with software, views of software technology, etc.)

Step 4 — Map Attributes to Existing Metrics

The last step is to map the relevant attributes to metrics that are being used in the organization. Remember that an attribute states "what" one wants to measure while the metrics defines "how" one measures something. The mapping consists of checking if the metrics are measuring the things (attributes) the users want to measure.

At this step, a GQM Structure is assembled for the user group. This structure shows the mapping between the user goals, the relevant entities, the relevant attributes, and the metrics used in the MF. This structure documents the data user group's needs measurement-wise.

At the end of this step one can derive a list of inconsistent, missing, and extraneous metrics from the user group point of view. Missing metrics are detected when a relevant attribute has no metric to measure it. Extraneous metrics are detected when a metric has no corresponding attribute in the GQM structure. Inconsistent metrics are detected when a metric used to measure a relevant attribute is not consistent with the user's goals. Typical consistency problems occur when: (1) the metric's scale or range of values is not suitable for the user needs; (2) the cost to apply a metric is unacceptable; or, (3) a metric cannot be applied when or where it is needed by the user group.

- Entry Criteria: Description of relevant attributes and user groups goals, and the description of the MF metrics are available.

- Input: Description of relevant attributes gathered in Step 3, description of the group goals gathered in Step 1, and description of the metrics and associated attributes gathered during the MF characterization phase.

- Procedure: Map the relevant attributes to the metrics that exist in the measurement framework. For each mapping, use the description of the goal's purpose to answer the following questions:

    - Is the metric's scale or range of values suitable for the user needs ?
    - Can the metric be applied when and where it is needed ?
    - Is the cost of applying the metric acceptable ?

  Establish a complete mapping from goals to metrics and identify:

    - Missing metrics: When a relevant attribute has no associated metric in the MF
    - Extraneous metrics: When a metric that apparently is useful to the user group has no mapping to a relevant attribute.
    - Inconsistent metrics: When a metric, in spite of measuring what the user wants, cause a negative answer to one of the questions listed before.

- Methodology: Establish the mapping between metrics and relevant attributes by comparing the description of relevant attributes with the attributes associated with the existing metrics. Interview data user group representatives to validate if the mapped metrics are consistent with the user goals.
- Output: A GQM structure for the user group (including goals, relevant entities and attributes, and existing metrics). A list of possible problems with the MF (missing metrics, inconsistent metrics, and extraneous metrics).
- Exit Criteria: The mapping from goals to metrics is complete and the possible problems have been identified.

3.3 Bottom-up Analysis

The data already collected by an organization is the most important asset of any MF. It is important for an organization to have means to explore its legacy data. Intelligent data exploration methods are an effective way of understanding and learning about the organization's business. This dissertation refers to them as a bottom-up methods, because the raw data is the starting point for better use and understanding of the data itself.

The top-down analyses are aimed at better planning and executing data collection. The bottom-up analyses are aimed at discovering new and useful information in the existing data, thus improving data awareness and data usage. The literature has many examples of the use of machine learning techniques to extract knowledge (new and useful information) from software engineering data sets [29, 32, 92, 101, 104, 107]. Our bottom-up analyses use Attribute Focusing [21] –

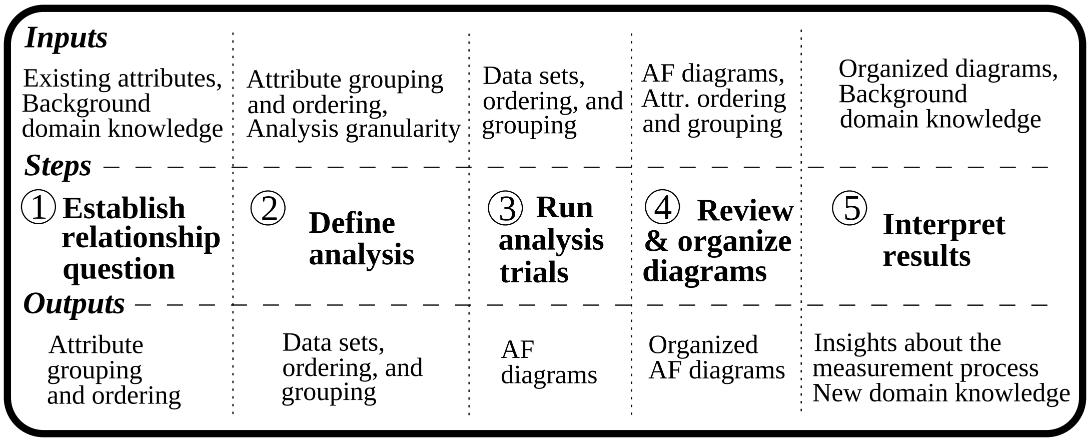

Figure 3.2: The AF-based Method

a data mining technique – to extract unexpected and useful information directly from the MF database.

3.3.1 The AF-based Method

The aim of the AF-based (bottom-up) method is to establish procedures to effectively apply the AF technique – maximizing knowledge discovery and minimizing discovery cost.

In the case of a measurement framework, the "experts" in the knowledge domain correspond to the MF data users and data managers. In this context, the bottom-up method allows the data users and managers to gain knowledge about: (1) their application domain (learn about the things they are measuring); and (2) the components of the measurement process (learn about the way they are measuring things).

In order to effectively apply the AF technique, the AF-based method goes through the five steps shown in Figure 3.2. In the first two steps the people in charge of applying the bottom-up method to the legacy data (i.e., data analysts) interact with the data users and managers to define the type of analysis that will be done. In the next two steps, the data analysts run the AF tool and organize the obtained results. In the last step, the results are reviewed by the data users. That's when knowledge discovery takes place.

Step 1 — Establish Relationship Questions

In the AF technique, it is very important to avoid the computation of uninteresting relations whenever possible. An uninteresting relation wastes machine time to compute, and yields uninteresting diagrams that will waste data user's time during the diagram reviews.

The AF Tool avoids uninteresting metric combinations based on user-defined data (metric) groups. The metrics grouped together are not correlated during the analysis. Our method uses generic relationship questions (GRQs) to select and group the data for the AF analyses. As explained in Section 2.4.4, each GRQ is used to state a set of relations that the data user wants to investigate empirically in an AF analysis.

The GRQs can be defined by: (1) interviewing user group representatives; or, (2) directly analyzing their measurement goals. In the latter case, it is very useful to have a GQM structure defined for the user group.

- Entry Criteria: Knowledge of the data available in the MF and understanding of what attributes they measure.
- Input: Data user domain knowledge and/or GQM structure and characterization of the MF components.
- Procedure: State a generic relationship question using Template 2.12 and answer the following questions:
  - What are the criteria used to determine the attribute classes ?
  - What attributes do compose each attribute class ?
  - What is the ordering of the classes (be sure to identify at least the explained class) ?
- Methodology: Interview the group representatives using the Template 2.12 described in Section 2.4.4, or use the group goals and description of available data to determine the GRQ.
- Output: GRQ, description of the attribute classes, and their ordering.
- Exit Criteria: The GRQ is completely described.

It is important to note that establishing a GRQ should be a very simple process. Attributes classes can easily be defined in any MF, and the ordering of the classes is directly determined by the empirical understanding of the cause-effect relationship between the involved attributes. If one is using GQM structures to define AF analyses, the nature of the relevant entities and the type of its attributes shall be used to define the attribute classes.

Step 2 — Define the Analysis

After establishing a GRQ for an AF analysis, the analysis itself must be defined. First, attributes identified in the GRQ must be mapped to the metrics in the MF. This is straightforward, if the information collected during the characterization phase is used.

Second, the data granularity and scope of analysis must be determined. Consider, for example, the GRQ 2.13:

> How do the "managerial constraints over the project" relate to the "logical features of the final products ?"
> Scope of the analysis: What products and projects should we consider ?
> Granularity: Should we analyze the data for the products individually or should we analyze classes of products ?

The scope and granularity should be directly derived from the user group goal and/or data use descriptions. The key here is to understand the group's purpose in using the data.

The data sets are extracted after the scope and granularity of the analysis are defined. This task is usually simple, but it may take a sizeable amount of effort if the data is not easily retrievable. The data sets may also need to be pre-processed and formatted to meet the data user and the AF tool data format requirements.

- Entry Criteria: GRQ is defined.
- Input: GRQ, and data and attribute description.
- Procedure: Identify the metrics associated with the attributes used in the GRQ. Determine what the scope and granularity of the analysis proposed by the GRQ is. Use the analysis scope to determine which data set will be used. Extract data set for analysis. Format data set according to the required granularity and the tool needs.
- Methodology: Interview the group representatives or use the group goals to determine the analysis scope and granularity. Use description of available data, metrics, and the analysis scope to extract the data. Use the description of the analysis granularity to pre-process the data.
- Output: Formated data set, and description of analysis scope and granularity.
- Exit criteria: the data is formated for the analysis.

Step 3 — Run the Analysis

The next step is to run the tool itself. This step is almost completely automated. The inputs are: (1) the metric groupings, (2) the maximum number of diagrams (relations) to be produced, (3) the interestingness cutoff level, and (4) the analysis dimension. The groupings are directly derived from the GRQs as previously discussed. The maximum number of produced diagrams is based on the time that the data users can spend looking at the diagrams. The interestingness cutoff determines the minimum interestingness value for which the tool will produce a diagram for a given relation. The higher this cutoff is, the more "interesting" (and less numerous) the produced diagrams are. The analysis dimension determines the maximum number of metrics that can appear in a diagram (e.g., a type three analysis results in up to 3-way diagrams).

- Entry Criteria: Formated data is available.
- Input: GRQ, and formated data.
- Procedure: Identify metric groupings, and determine number of diagrams, the analysis cutoff and dimension. Import formated data into the tool and input the previous parameters. Run the tool.
- Methodology: Use GRQ to identify metric groups (each attribute class corresponds to a group). Manipulate the cutoff and maximum number of diagrams to obtain a reasonable number of "interesting" diagrams. Set the analysis dimension based on the number of attribute classes.
- Output: AF Diagrams.
- Exit criteria: The AF tool has finished its analysis.

Step 4 — Organize the Diagrams

Although many uninteresting diagrams have already been pruned away with the metric groupings, there may still be diagrams that are unsuitable for the data user's review. The next step is to manually review the diagrams before they are shown to the data users. It may be necessary to (re-)run the analysis trials if: (1) too few diagrams were found for a given cutoff; or (2) missing or skewed data is driving the discoveries.

After a sizable number of useful diagrams have been compiled, they are organized to facilitate the data user's inspection. Diagrams may be grouped in several ways. These "groups of diagrams" allows the data users to concentrate on unique reasoning threads while looking at them. They may also be used to produce more complete summaries of relations between explanatory and explained variables (see Table 4.10 for an example of such a summary).

- Entry Criteria: AF diagrams are available.
- Input: AF diagrams, GRQ, and description of the analysis granularity and scope.

- Procedure: Eliminate diagrams that show obvious facts or were produced because missing or skewed data drove the "discovery". Return to Step 2 if too few "useful" diagrams were produced. Otherwise, organize produced diagram in groups. If useful information can be drawn by comparing different diagram groups, consider producing tables to summarize the results.

- Methodology: Review available diagrams one by one. Discard useless diagrams and organize the others using some consistent criteria (e.g, group diagrams with the same explanatory metrics and related explained metrics together).

- Output: Organized AF Diagrams and information summary.

- Exit Criteria: The information produced by the AF Tool is organized.

The following algorithm may be used to group diagrams. This algorithm puts diagrams with the same explanatory metrics and related explained metrics in the same group:

1. Organize all the $\mathcal{N}$ diagrams obtained from the AF tool by order of interestingness.

2. Discard diagrams that are clearly uninteresting.

3. Following the order of interestingness given by the AF Tool, select the first diagram.

4. Select all other diagrams that have the same explained metric, and an explanatory metrics in the same groups as the explanatory metrics of the diagram selected in step 3.

5. Group all the diagrams selected in step 3 and 4 by order of interestingness in a unique "explanatory group" to be shown together to the data users.

6. Remove the diagrams gathered in step 4 from the overall group of produced diagrams and return to step 3, if there still are diagrams from the original $\mathcal{N}$ diagrams produced.

If one wants diagram groups with the same 'explanatory metrics' and related 'explained metrics', he/she can modify step 4 in the above algorithm as follows:

4. Select all other diagrams that have the same set of explanatory metrics and an explained metric in the same group as the explained metric of the diagram selected in step 3.

Step 5 — Review the Diagrams

The last step of the AF-based method is the analysis of diagram groups by the data users. The diagram groups have many types of information in them:

1. Unexpected correlations between metrics (direct analysis of a N-way diagram).

2. Unexpected value distributions (direct analysis of a 1-way diagram).

3. Unexpected (in)consistencies in the relationships between explanatory metrics and related explained metrics (direct from analysis of a diagram group or tables with summarized results).

New knowledge is gained when the data users apply their background knowledge to interpret the information contained in the diagrams. There are two types of domain knowledge to be gained in this way: (1) insights into their application domain; and (2) insights about the components of the measurement process.

The first type of result is what is traditionally expected from the AF technique. The technique helps the experts to gain new insights into their activities. These insights may lead the data users to take adaptive, corrective or preventive actions to improve the way they do business.

The second type of result happens when the AF diagrams lead the data users to realize that some previous assumption about the data or measurement process is incorrect. This may lead them to modify their measurement goals, metrics, predictive models, and data collection procedures.

The diagrams might also have interesting information that raises new questions about the data behavior. This usually happens when the new information cannot be easily interpreted (transformed in knowledge) by the expert. In this case, the new questions may be used to define new data analyses, if the data needed to answer them are already available in the MF. When the MF does not have the needed data, these questions may be transformed in new measurement goals and fed back to the GQM-based method.

- Entry Criteria: The information produced by the AF Tool is organized.

- Input: Organized AF Diagrams and result summaries.

- Procedure: Review diagrams (in groups) and result summaries. Document any new insight. Record the background knowledge used to gain the insight and the new knowledge gained with the insight. Return to Step 1 of the AF-based method, if an insight raises a new question about the data.

- Methodology: Review diagrams together with data users or managers.

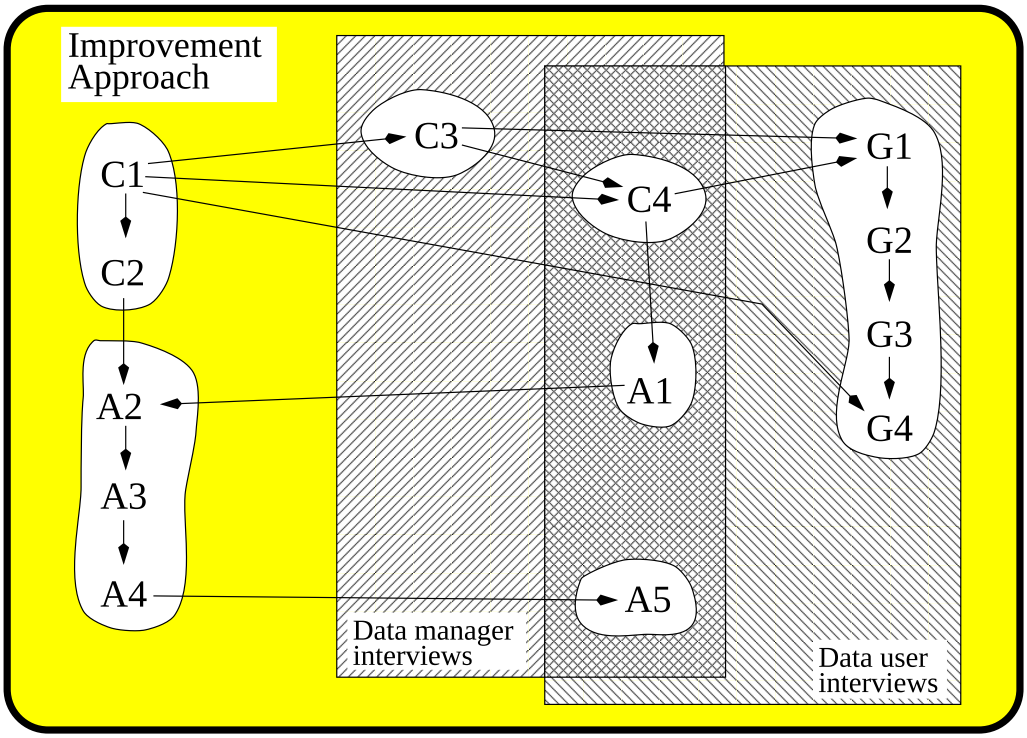

Figure 3.3: Dependencies Between the Approach Phases and Steps

- Output: New knowledge.
- Exit Criteria: All diagrams were reviewed.

3.4 Overview of the Whole Approach

Figure 3.3 shows the dependencies between the three phases of the approach. Each phase is represented by its steps:

- Characterization phase

  C1- Metrics characterization

  C2- Data characterization

  C3- Data uses and user groups characterization

  C4- Attributes characterization

- Bottom-up analysis phase (AF-based method)

  A1- Establish generic relationship question (GRQ)

A2- Define analysis

A3- Run analysis trials

A4- Review and organize diagrams

A5- Interpret results

- Top-down analysis phase (GQM-based method)

G1- Capture data user goals

G2- Identify entities

G3- Identify attributes

G4- Map attributes to metrics

Arrows are used to indicate the dependency between the steps in Figure 3.3. They indicate that the characterization phase is a pre-condition to execute top-down and bottom-up analyses. The arrow between C4 and A1 indicates that a good understanding of the MF attributes is a pre-condition to define generic relationship questions (GRQs) for the AF analyses. The arrow between C2 and A2 indicates that a good understanding of the data is needed to define an AF analysis. The arrows from C3 and C4 to G1 indicates that a good understanding of user groups, data uses, and attributes is needed to capture the data user goals. The arrow between C1 and G4 indicates that a good understanding of the MF metrics is necessary to do the mapping between relevant attributes and existing metrics.

The white shapes in Figure 3.3 indicate the steps that should be executed together or interactively. For example, the four steps of the GQM-based method should be executed in one interactive interview with the data users. AF analyses definition, execution, and diagram review (A2, A3, and A4) should also be executed together and interactively. The same is true for metrics and data characterization (C1 and C2).

The rectangles in Figure 3.3 show what steps are done in interviews with data managers and data users. The intersection between the rectangles indicates that the steps C4 (characterizing attributes), A1 (establishing GRQs), and A5 (interpreting AF results) can be done in interviews with data managers or data users. Step C3 (characterizing data uses and user groups) is done in interviews with the data managers, and the GQM-based method is applied interviews with the data users.

Although there is no dependency between the AF and GQM-based methods, they can interact with each other. Figure 3.4 shows this interaction. GQM structures can be used to derive generic relationship questions (A1) and define AF analyses (A2). The AF analyses may produce interesting facts that raise new

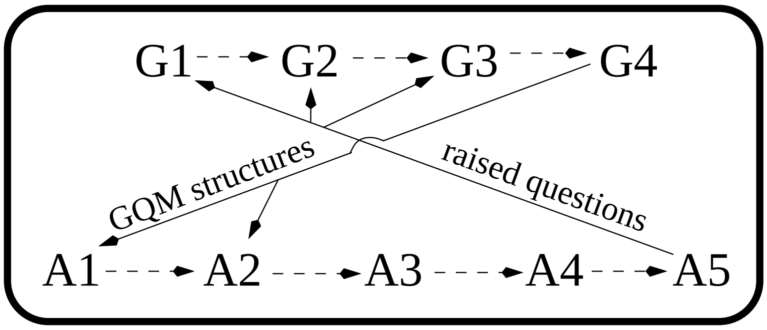

Figure 3.4: Interaction Between the AF and GQM-based Methods

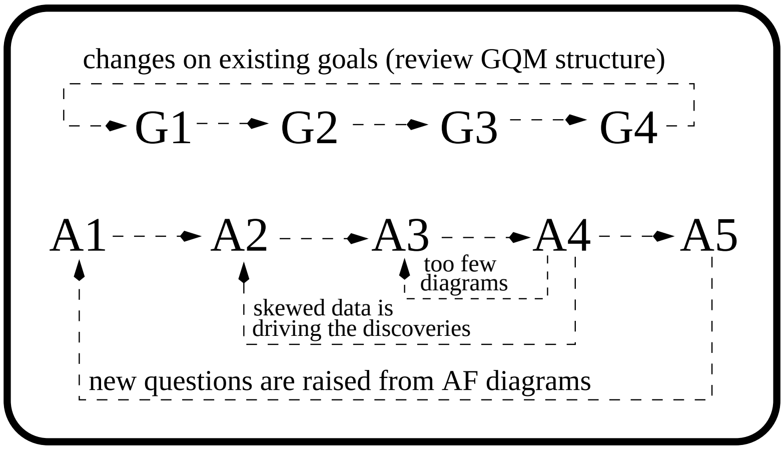

Figure 3.5: Iterating the AF and GQM-based Methods

questions about the data behavior. If those questions are important and cannot be answered by the existing data, they should be used to define new measurement goals (G1), relevant entities (G2), and relevant attributes (G3) for the MF.

Figure 3.5 shows that both the AF and GQM-based methods are iterative. GQM structures should be revised if there is a change on the data user goals. In this case, the existing GQM structure can be used as input to a new cycle of the GQM-based method.

The iterations are more fine grained in the AF-based method. The analysis must be repeated if too few diagrams are produced. In this case, the data analyst can re-run the AF tool (A3) resetting some of the analysis parameters (e.g., by lowering the cutoff value). The review of the diagrams (A4) may also reveal that a problem in the data is driving the discoveries. In this case, the analysis needs to be redefined (A2). Last but now least, interesting facts may lead to new questions about the data behavior. These questions may originate new AF analyses if the needed data is available in the MF.

Chapter 4

Case Study

This chapter describes the experience of applying our approach to the IBM Toronto Laboratory's Customer Satisfaction (CUSTSAT) measurement framework. The CUSTSAT data is collected annually by surveys carried out by an independent party. Its purpose is to evaluate customer satisfaction with products of IBM's Software Solutions Division (SWS) and their competitors. The IBM Toronto Laboratory (Toronto Lab) is only one of the several IBM Software Solutions laboratories that use the CUSTSAT data. Inside the IBM Toronto Laboratory, the CUSTSAT data is used by several different groups (e.g., development, service, support, and senior management).

IBM surveys a large number of customers from several different countries. All the data is stored in one database. Currently, this database already stores several years of CUSTSAT data.

The large amount of data and the diversity of groups that are interested in this data made it desirable to apply our approach to the CUSTSAT MF. Our two main objectives in doing that were: (1) better understanding of the user groups' needs with respect to the CUSTSAT measurement; and (2) better exploration of the data already stored in the CUSTSAT database.

We effectively started this work in the summer of 1995. Most of CUSTSAT MF characterization and some AF analyses were done that year. In 1996, the MF characterization was updated to the values reported in Section 4.1, the AF analyses reported in Section 4.3 were run, and the top-down analysis reported in Section 4.2.1 was executed. In 1997, the top-down analyses reported in Sections 4.2.2 and 4.2.3 were executed.

4.1 Characterization of the CUSTSAT MF

The first step of our approach was to document the metrics that composed the CUSTSAT MF, their user groups, and how these groups used this data. We focused on the data related to the Toronto Laboratory products. We did not

Attribute Groups	Description
Company information	Company address, name, primary business, etc.
Contact information	Name, phone number, title, job, etc.
Product information	Name, vendor, operating system, version, etc.
CUPRIMDS / O Sat	Customer satisfaction with main product attributes (see Table 4.2).
CUPRIMDS Importance	Relative importance of main product attributes.
Documentation Sat	Satisfaction with documentation attributes.
Multi-cultural factors	Sat. w/multi-cultural factors like translations, support for international characters, etc.
Product distribution	Satisfaction with product delivery.
Ability to acquire product	Sat. with ability to acquire product.
Technical service	Sat. with developer technical service and support.
Local product support	Sat. with local support, education, and sales.
Disposition to upgrade	Disposition to recommend, upgrade, and re-purchase the product.
Price Sat	Sat. w/product price, terms, and conditions.
Communications	Questions to evaluate the marketing communication channels.
Sale Channel	Questions to evaluate sale channels.
Value system	Questions for market value system evaluation.
Topology	Questions to evaluate customer computing environment.
Decision Maker	Questions to identify purchase decision makers.

Table 4.1: Main Attribute Groups

work with any metrics or user groups associated with products developed in other IBM laboratories.

The information collected in this step was gathered from existing documents in the MF (e.g., the survey questionnaire and MF database schema) or by interviewing the data manager responsible for the CUSTSAT data in the Toronto Laboratory. We have used forms to document the information collected about metrics, attributes, user groups, and data uses. These forms are shown in Appendix A. They were also used to check for the completeness of the information collected during the interviews with the Toronto Lab data manager.

Attribute	Description
C_SAT	Satisfaction with product capability (functionality)
U_SAT	Satisfaction with product usability
P_SAT	Satisfaction with product performance
R_SAT	Satisfaction with product reliability
I_SAT	Satisfaction with product installability
M_SAT	Satisfaction with product maintainability
D_SAT	Satisfaction with product documentation
S_SAT	Satisfaction with product technical service and support
O_SAT	Overall satisfaction with the product

Table 4.2: CUPRIMDS/O Sat. Attributes

4.1.1 Metrics and Attributes

The task of identifying what metrics compose the CUSTSAT framework was simple. Most of the metrics corresponded to questions in the survey questionnaire (SQs). The exception is the customer contact information stored separately in the CUSTSAT database. Next, the metrics' meanings were recorded. This corresponds to the attribute that a metric is supposed to be measuring. This task was facilitated by the fact that we were working with a questionnaire that was geared toward the customers. The formulation of a questionnaire question explains what it wants to measure. Terms like "capability," "performance," or "maintainability" are explained when they are used in the questionnaire. This eliminated the need of interviewing data users to record the metrics' meaning (step 4 of the MC process).

Overall, we identified more than 100 attributes that are measured (mostly in ordinal scale) in the CUSTSAT framework. In order to preserve IBM proprietary information, the complete list of metrics and attributes used in the CUSTSAT MF is not listed in this dissertation. Nonetheless, Table 4.1 lists the main attribute groups that are measured by the CUSTSAT MF, and Table 4.2 describes the most important of these groups.

4.1.2 Available Data

The CUSTSAT data is collected by phone surveys since 1993. The survey process lasts for several months and the survey questionnaire is modified annually. The survey covers all products of IBM Software Solution (SWS) Division and their competition. Each interview correspond to one data point.

User Groups	Description
SWS Div	Software solutions division headquarters
Toronto Lab	Toronto laboratory senior management
DB Mgmt	Database technology management
DB Product Dev.	Database products development teams
DB Common Tech.	Database common technology development
DB Customer Sup.	Database customer service support
ID DB	Database information development
DB Usability Grp	Database products usability team
AD Mgmt	Applications development management
AD Dev	Development groups for AD tools, Fortran, C/C++ (including VisualAge), 390 languages, and AS/400 languages
AD Support	Service and support for AD languages
ID AD	Information development for AD tools, Fortran, C/C++ , 390 languages, and AS/400 languages
MKT Plng	Market planning
Pricers	Product pricers
Country Sales	IBM Canada country sales management group
Marketing	Software marketing and services
ISM	IBM software manufacturing (Boulder CO)

Table 4.3: User Groups

30 to 600 data points are collected for each IBM product per year. In 1996, a total of 13,000 data points were collected in North America. Smaller samples are also collected in Europe and Japan. All the data is entered in a unique database called Customer Information System (CIS). The CIS is physically located at Toronto, but can be accessed from any SWS laboratory. In 1997, an intranet Web-based interface for CIS was created. This interface produces automatically most of the standard reports and data analysis done with the CUSTSAT data. Any other type of access to CIS has to be requested to and granted by one of the CUSTSAT data managers.

4.1.3 Data Uses and User Groups

The data uses and user groups were characterized in interviews with the data manager. The data manager was asked to describe all analyses and presentations done with the CUSTSAT data. Each type of data analysis or presentation (DA/P) corresponded to a distinct data use. The data use descriptions included the frequency with which the DA/Ps were performed, the list of metrics used in them, the granularity and scope of the DA/Ps, and the list of people that were interested in the DA/Ps. A list of user groups was compiled by mapping the list of people that used the DA/Ps to the formal groups inside the laboratory. The data manager was asked to describe the listed user groups. The user group descriptions included: (1) a statement of the data manager's perception of the group's objectives in using the data; (2) a list of the data uses associated with the group; and (3) a subjective ranking of the importance of the CUSTSAT data to the group.

Overall, we have identified 17 user groups that can be divided into four major areas: senior management, database development, compiler development, and support (e.g. market analysis, marketing, and sales). The user groups are listed in Table 4.3. We also identified about 16 different DA/Ps (data uses) associated with the CUSTSAT data. The CUSTSAT data uses (related to the Toronto laboratory software products) are listed below:

- Division level analysis - high level evaluation of the division products by the management. This includes review of CUPRIMDS/O satisfaction, price satisfaction, and disposition to upgrade, re-purchase, recommend the product. AD and DB products are analyzed as whole against the competition.

- Market plan for AD - CUPRIMDS/O satisfaction data is used by the laboratory senior management during market planning for application development. The data is considered as a whole, against the competition, for all AD products by market place and/or platform type.

- Satisfaction check for product release - CUPRIMDS/O satisfaction data is used to certify the product quality before a new official release. It uses the customer satisfaction data gathered for the product beta version.

- CUPRIMDS/O SATISFACTION review of DB products - DB personnel periodically review comparisons of their products against the competition. The review usually covers CUPRIMDS/O satisfaction and importance attributes.

- Annual DB CUSTSAT review - DB personnel also does a round up annual review of their products attributes against the competition. The review usually covers CUPRIMDS/O satisfaction and importance, disposition to upgrade, re-purchase, recommend products, and O_SAT$\times$CUPRIMDS correlation.

- Investigate customer D_SAT - when D_SAT is low for a given DB product, the information development personnel try to characterize the documentation problems using the customer comments on the documentation. Sometimes, they also call dissatisfied customers.

- Investigate customer U_SAT - when usability satisfaction is low for a given DB product, the usability group tries to characterize problems with the product user interface using the customer comments on usability.
- Investigate customer S_SAT - the DB technical service support group try to review the comments of all customers dissatisfied with their service. They also call them whenever it is possible.
- Product market evaluation - the market planning group wants to use the value system, topology, and decision maker attributes to evaluate if the products are effectively being sold to the markets they were planned for.
- Customer information management - the CIS system is used to store and retrieve company and contacts information to be used in different market surveys (including the CUSTSAT survey). On the flip side, the surveys are used to update and expand the CIS contact list.
- Price evaluation - used by the Pricers to review the customer satisfaction with the price, terms, and conditions associated with the lab products.
- Evaluation of marketing communication channels - the marketing personnel use the communication channels characterization to evaluate marketing communication strategies.
- Sale channels characterization - the marketing personnel uses the characterization of the sale channels to improve marketing strategies.
- Price satisfaction characterization - the marketing personnel uses the information on customer satisfaction with price, terms, and conditions to improve marketing strategies.
- Local support evaluation - Canada country sales group uses the local support satisfaction data to evaluate the customer satisfaction with the local sales offices.
- Software distribution evaluation - the IBM Software Manufacturing (ISM) analyzes the customer satisfaction with product distribution to evaluate and improve their services.

Table 4.4 associates data uses with the user groups. Its third column shows when the data is used by the user groups. Table 4.5 associates data uses with attribute groups. Its third column shows the granularity at which the users look at the data. This dissertation will not give more detailed descriptions of user groups and data uses as this information is considered IBM proprietary.

Data Uses	User Groups	When
Division level analysis	SWS Div	yearly
Market plan for AD	SWS Div and Toronto Lab	yearly
Satisfaction check for product release	Toronto Lab, DB or AD Mgmt, and DB or AD product development group	when a new product is being released
CUPRIMDS/O SAT review of DB products	DB Mgmt., DB Common Tech., DB Product Dev., DB Customer Support, ID DB, and DB Usability Grp.	monthly
Annual DB CUSTSAT review	DB Mgmt., DB Common Tech., DB Product Dev., DB Customer Support, ID DB, and DB Usability Grp.	yearly
Investigate customer D_SAT	ID DB	when D_SAT is low for a DB product
Investigate customer U_SAT	DB Usability Grp.	when U_SAT is low for a DB product
Investigate customer S_SAT	DB Customer Support	when there is a customer with low S_SAT
Product market evaluation	MKT Plng	(planned) yearly
Customer information management	MKT Plng	during survey planning
Price evaluation	Pricers	(planned) yearly
Evaluation of – marketing – communication channels	Marketing	yearly
Sale channels characterization	Marketing	yearly
Price satisfaction charac.	Marketing	yearly
Local support evaluation	Country Sales (Canada)	Not available
Software distribution eval.	ISM	monthly

Table 4.4: Data Uses × User Groups

Data Uses	Attribute Groups	Granularity
Division level analysis	Price Sat., CUPRIMDS/O Sat, disposition to upgrade	product class by platform type by year
Market plan for AD	CUPRIMDS/O Sat	market place by platform type by year
Satisfaction check for product release	CUPRIMDS/O Sat	by product release beta test data
CUPRIMDS/O Sat review of DB products	CUPRIMDS/O Sat and Imp;	product release by year to date
Annual DB CUSTSAT review	CUPRIMDS/O Sat and Imp; disposition to upgrade	product release by year
Investigate customer D_SAT	Documentation satisfaction	product rel. by month
Investigate customer U_SAT	U_SAT and suggested improvements	product rel. by month
Investigate customer S_SAT	Sat. with manufacturer technical service support	DB products by month
Product market evaluation	Value system, topology, and decision maker	by product release
Customer information mgmt.	Contact, company, and product basic information	by product or product class
Price evaluation	Sat. w/price, terms, and conditions	by product release
Evaluation of marketing communication channels	Marketing comm. channels	by product
Sale channels characterization	Sale channels	by product
Price satisfaction charac.	Sat. w/price, terms, and conditions	by product
Local support evaluation	Satisfaction w/local support	product by country
Software distribution eval.	Sat. w/product distribution	products from all labs by month

Table 4.5: Data Uses $\times$ Attribute Groups

4.2 Top-down Analyses in the CUSTSAT MF

We applied our GQM-based method to a limited number of data user groups in order to test the method feasibility and effectiveness. We built GQM structures for three user groups to propose improvements in the CUSTSAT questionnaire based on the obtained results. The three chosen groups are associated with the database product development at the laboratory:

1. the DB customer service and support group.
2. the DB information development (documentation) group.
3. the DB usability group.

We used structured interviews [78] to build GQM structures for these groups. We interviewed a senior representative of each group. All the material for the interview was prepared beforehand. It included:

- a complete list and description of the metrics and DA/Ps associated with the group.
- a tentative description of our perception of their goals.
- a tentative list of entities and attributes that we believe were relevant for them.
- a complete list of questions and topics to be discussed during the interview.

This material was prepared based on the MF documents and the group internal documents. All the material was integrated in a single interview script. The script used during the documentation group interview is shown in Appendix B. Similar scripts were written to interview the other groups.

The scripts try to capture: (1) the group goals in using the CUSTSAT data; (2) the relevant entities associated with their work; (3) the relevant attributes they want to measure through the CUSTSAT survey; (4) and the metrics (questionnaire questions) that are effectively measuring them. The interviews were also used to validate and rate the importance of the DA/Ps and metrics associated with each data user group. These last activities can be considered part of the MF characterization phase.

4.2.1 Service Support Interview

The Service Support interview was done in two meetings in 1996. The first step of the interview was to ask the interviewee for comments on the data analyses and presentations (DA/Ps) done for the group. This step had two objectives: (1) motivate and focus the rest of the interview around the CUSTSAT MF; (2) validate our understanding of their data usage (including assessing the importance of the data for them).

The second step was to capture their goals in using the CUSTSAT data. This part was supported by the previous discussion of the group data usage. We asked the interviewee to describe what the group wanted to achieve in using the CUSTSAT data, and expressed it in the form of GQM goals. We captured the following goals:

Goal 1: Analyze the service support process in order to characterize its key areas with respect to customer satisfaction and dissatisfaction.

Goal 2: Analyze the customers in order to understand them with respect to expectations for support service.

Goal 3: Analyze the service support areas with which the customers are dissatisfied in order to improve them with respect to customer satisfaction.

The next step was to discuss the relevant entities, attributes, and metrics associated with those goals. We started by identifying the relevant entities. From the entities and goals, we have discussed the relevance of the following attributes:

Entity 1: Service support (SS) process

- Attribute 1.1: Overall customer satisfaction with SS
- Attribute 1.2: Improvements suggested to the SS by the customer
- Attribute 1.3: Customer satisfaction with time to resolution
- Attribute 1.4: Customer satisfaction with SS responsiveness
- Attribute 1.5: Aspects customer liked most about the SS
- Attribute 1.6: Aspects customer disliked most about the SS
- Attribute 1.7: Customer satisfaction with the SS commitment level

Entity 2: Support team

- Attribute 2.1: Customer satisfaction with support team skill and knowledge

Entity 3: Solution/resolution provided

- Attribute 3.1: Quality of the solution
- Attribute 3.2: Satisfaction with the problem resolution
- Attribute 3.3: Degree to which resolution met expectations

Attribute 3.4: Reasons why resolution did not meet expectations

Entity 4: Reported problem

Attribute 4.1: Severity
Attribute 4.2: Type

Entity 5: Customer contact (surveyed person)

Attribute 5.1: Role in organization (job responsibilities)

Entity 6: Customer organization

Attribute 6.1: Primary business
Attribute 6.2: Type of activities

Entity 7: Product being supported

Attribute 7.1: Product name
Attribute 7.2: Product version
Attribute 7.3: Date the product was installed in the organization

The above list includes the attributes associated with existing metrics as well as new attributes suggested by the interviewee. In the case of new attributes, it was important to make sure that we understood and recorded their meaning. Let us consider "attribute 1.7" as an example. According to the interviewee, this attribute refers to the degree to which the service support meets the commitment level contracted by the customer. Those levels are established in the support contract and correspond to well-defined preconditions on the time that IBM should take to provide a satisfactory solution to customer problems.

Figure 4.1 depicts the GQM structure for service support group. It shows the mapping from the attributes to the metrics (questions in the survey questionnaire). In the structure, the metrics are referred to by the question number in the survey questionnaire. The rectangles indicate that the attribute was suggested by the interviewee's goals but is not being measured yet. From Figure 4.1, we concluded that there are eight missing metrics from the service support point of view (rectangles). These metrics are needed to measure attributes: 1.5, 1.6, 1.7, 3.2, 3.3, 3.4, 4.1, and 4.2. However, attributes 4.1 and 4.2 (open rectangles) were considered too difficult to measure[1]. The crossed out metrics – Q45c, Q45a, Q45b, and Q6a – indicate that their associated attributes were considered irrelevant by the interviewee. They are extraneous from the service support point of view.

During the interview, we also checked if the interviewee had any comments on the structure of the metrics. In particular, we have asked for comments on the wording and the ranges of values of the questions. The interviewee comments and GQM structure for the service support group were recorded. They were used as input to the annual questionnaire review, and will be used in future improvement cycles with the SS group.

[1]The surveys can be taken up to six months after the problem occurred. It might be difficult for the customer to classify the severity and type of problems in those cases.

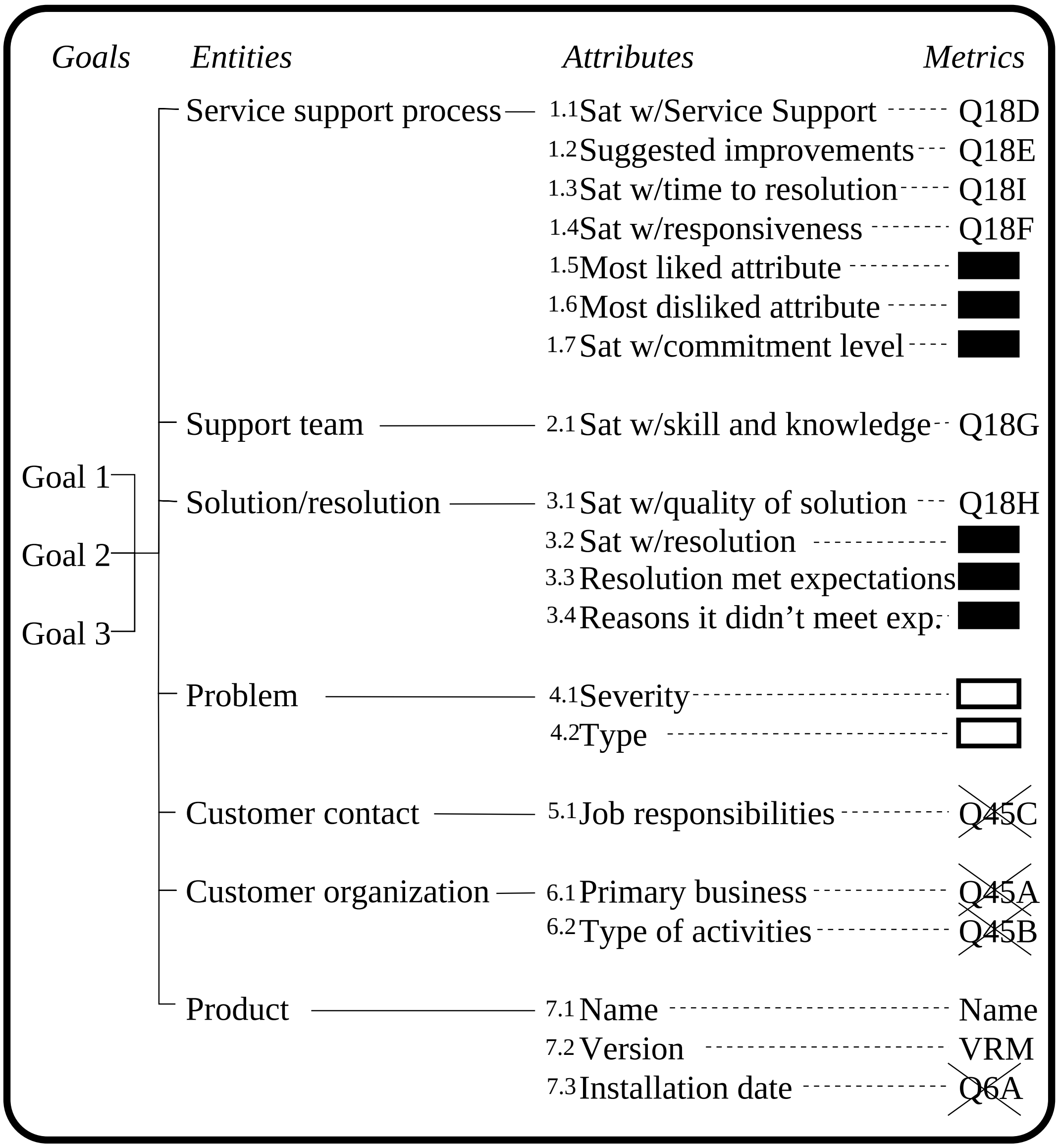

Figure 4.1: GQM Structure for the Service Support Group

4.2.2 Information Development Interview

The interview with the information development (documentation) group was done in two rounds in 1997. This interview followed a script very similar to the service support interview. The first step of the documentation group interview was to ask the interviewee for comments on the data analyses and presentations (DA/Ps) done for the group. The second step was to capture their goals in using the CUSTSAT data. We captured the following goals:

Goal 1: Analyze the documentation deliverables in order to characterize their key areas with respect to customer satisfaction and dissatisfaction.

Goal 2: Analyze the documentation deliverables with which the customers are dissatisfied in order to improve them with respect to customer satisfaction.

Goal 3: Analyze the documentation deliverables in order to understand them with respect to their relative importance to the customer.

The next step was to discuss the relevant entities, attributes, and metrics associated with those goals. We started by identifying the relevant entities. From the entities and goals, we have discussed the relevance of the following attributes:

Entity 1: Documentation provided by the vendor

- Attr. 1.1: Customer satisfaction with documentation
- Attr. 1.2: Customer rating of documentation importance
- Attr. 1.3: Types of documentation deliverables used by customer
- Attr. 1.4: Most important deliverable for the customer
- Attr. 1.5: Aspects customer liked most about the documentation

Entity 1.1: Printed manuals

- Attr. 1.1.1: Customer satisfaction with printed manuals
- Attr. 1.1.2: Improvements suggested to the printed manuals by the customer
- Attr. 1.1.3: Most used printed manuals
- Attr. 1.1.4: Whether or not the customer uses manuals in printable format

Entity 1.2: On-line help screens

- Attr. 1.2.1: Customer satisfaction with on-line help screens
- Attr. 1.2.2: Improvements suggested to the on-line help screens by the customer

Entity 1.3: Soft-copy books

- Attr. 1.3.1: Customer satisfaction with soft-copy books

Attr. 1.3.2: Improvements suggested to the soft-copy books by the customer

Entity 1.4: Tutorials

Attr. 1.4.1: Customer satisfaction with tutorials

Attr. 1.4.2: Improvements suggested to the tutorials by the customer

Entity 1.5: Other type of document

Attr. 1.5.1: Name

Entity 2: Translation Process

Attribute 2.1: Customer satisfaction with translations

Attribute 2.2: Improvements suggested to the translation by the customer

Attribute 2.3: Customer ratings of the importance of having documentation translated to their native language

Entity 3: Customer contact (surveyed person)

Attribute 3.1: Role in organization (job responsibilities)

Entity 4: Customer organization

Attribute 4.1: Primary business

Attribute 4.2: Type of activities

Entity 5: Product being supported

Attribute 5.1: Product name

Attribute 5.2: Product version

Attribute 5.3: Platform (operating system)

The above list includes the attributes associated with existing metrics as well as new attributes suggested by the interviewee. Figure 4.2 depicts the GQM structure for documentation group. As before, the metrics are referred to by the question number in the survey questionnaire. The rectangles indicate that the attribute was suggested by the interviewee's goals but is not being measured yet. These missing metrics are needed to measure attributes 1.4 , 1.5, 1.1.3, 1.1.4, and 1.5.1. Attribute 1.1.3 is considered difficult to measure because the manual names are product specific.

The crossed out metrics – Q15h, Q15i, Q14b, Q14c, Q20m, Q45c, Q45a, and Q45b – indicate that their associated attributes were considered irrelevant by the documentation group. They are extraneous from their point of view. It is worth noticing that the tutorials questions – Q15h and Q15i – was being used only by this group and are now considered extraneous overall. They will certainly be removed from next year survey questionnaire.

As before, the interviewee comments and GQM structure for the group were recorded. They will be used as input to the annual questionnaire review (modification), and in future improvement cycles with the documentation group.

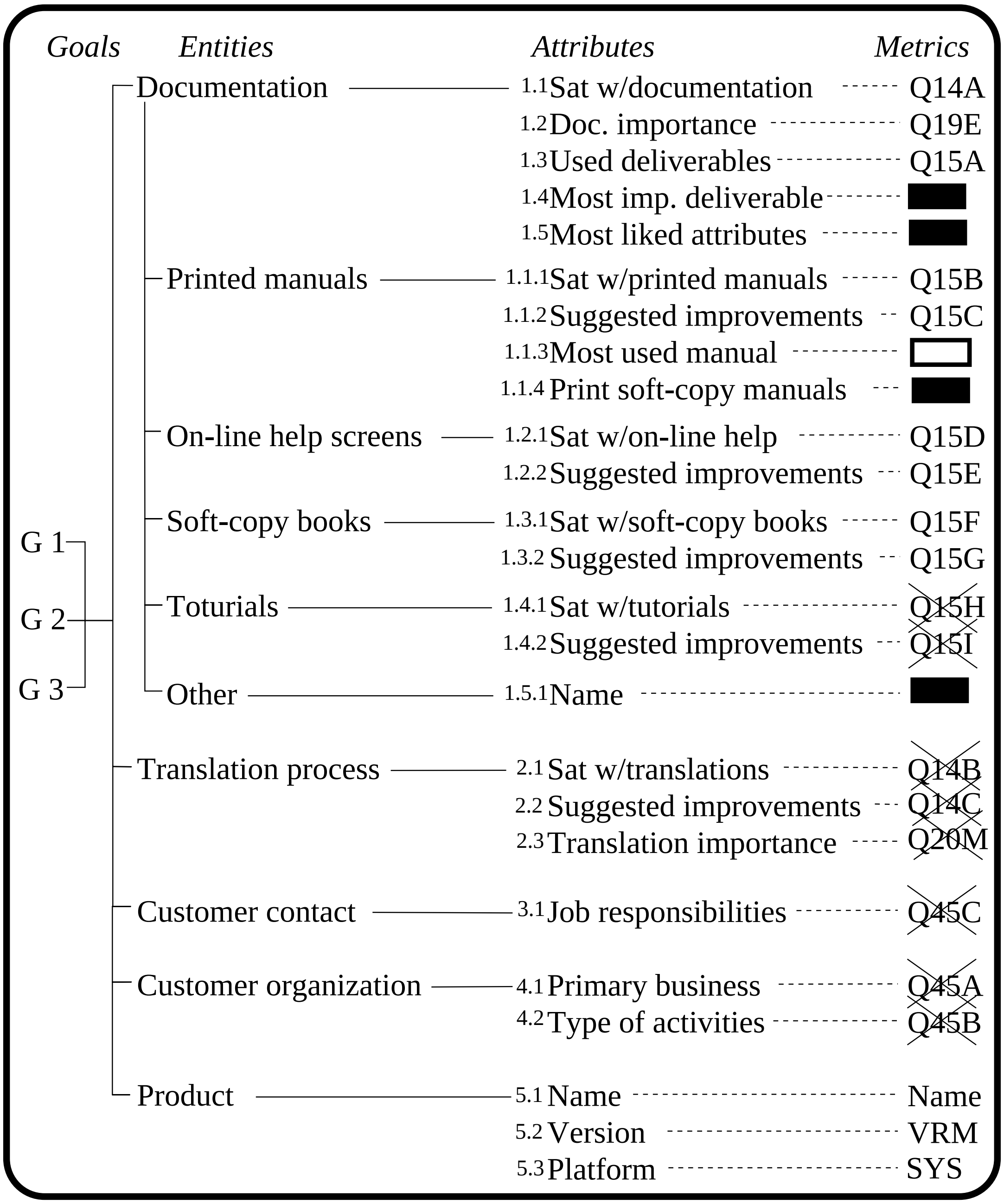

Figure 4.2: GQM Structure for the Documentation Group

4.2.3 Usability Interview

The usability interview was also done in 1997. It followed a script quite different from the previous ones. The usability group does not use the survey data as much as the other groups do. They rely on a more specific survey done through the World Wide Web with users of beta versions their products. This survey is product specific and focuses on usability, installability, and the overall product offering (the usability group is interested in all aspects of the user's "product-experience"). In the first contact with the group, the group manager suggested that we talk with the person responsible for those beta-surveys inside the group. Contrary to the other interviewees this person was not a manager but an expert in measurement.

The GQM process could not be completely applied with this interviewee. He could list the group goals and identify a large set of usability attributes, but he could not identify which of those attributes should be surveyed by the CUSTSAT MF. One of the reasons for this is that the main usability attributes are still being discussed at IBM. The following measurement goals were identified during this interview:

Goal 1: Analyze the customers (users) in order to characterize them with respect to expertise (in usage) and familiarity with the competition.

Goal 2: Analyze the product in order to characterize it with respect to customer acceptance.

Goal 3: Analyze the user interface in order to evaluate it with respect to intuitiveness, visual appeal, and task efficiency.

Goal 4: Analyze the full product offering in order understand its problems with respect to usability, reliability, capability, and installability.

During this interview we learned that there is a committee at the corporate level that is currently defining a standard set of software usability attributes to be measured during the software development process. We contacted the representative of this group inside the Toronto Laboratory. He was able to suggest a list of relevant candidate attributes for the CUSTSAT survey. The following relevant entities and attributes were identified:

Entity 1: Full product offering

- Attr. 1.1: Overall customer satisfaction
- Attr. 1.2: Reasons that the customer has to be dissatisfied with the product

Attr. 1.3: Satisfaction with first experience with the product
Attr. 1.4: Satisfaction with integration of product parts
Attr. 1.5: Satisfaction with the product capability (functionality)
Attr. 1.6: Improvements suggested to the product functionality by the customer
Attr. 1.7: Satisfaction with the product reliability and availability
Attr. 1.8: Improvements requested in reliability by the customer
Attr. 1.9: Satisfaction with the product documentation
Attr. 1.10: Product name.
Attr. 1.11: Product version.
Attr. 1.12: Product platform (operating system).

Entity 1.1: User interface

- Attr. 1.1.1: Overall satisfaction with the product ease of use
- Attr. 1.1.2: Improvements suggested to the user interface by the customer
- Attr. 1.1.3: Satisfaction with user model (user model reflects the way the customer works)
- Attr. 1.1.4: Satisfaction with ease of interaction and navigation (ease of moving around and do things using input devices – e.g., mouse and keyboard)
- Attr. 1.1.5: Satisfaction with user interface consistency (similar features look similar in different parts of the product)
- Attr. 1.1.6: Satisfaction with video and audio appeal (appearance is pleasant in terms of color, layout, and graphics; the use of sound enhances the product)
- Attr. 1.1.7: Satisfaction with task execution efficiency (tasks can be completed in the minimal number of steps).
- Attr. 1.1.8: User assistance satisfaction (features that provide help are useful and easy to use).

Entity 1.2: Installation process

- Attr. 1.2.1: Satisfaction with product installability
- Attr. 1.2.2: Satisfaction with product uninstallability
- Attr. 1.2.3: Improvements suggested to the installation process by the customer

Entity 2: User (customer contact)

Attribute 2.1: Familiarity with competition
Attribute 2.2: Expertise with product

The GQM structure for the usability group is shown in Figure 4.3. This figure includes the attributes associated with existing metrics as well as new attributes suggested by the second interviewee. As before, the metrics are referred to by the question number in the survey questionnaire. The rectangles indicate that the attribute was suggested by the interviewee but is not being measured yet. These missing metrics are needed to measure attributes 1.3, 1.4, 1.1.3-1.1.7, 1.2.3 and 2.2. Attribute 2.2 (user expertise) is considered difficult to be measured

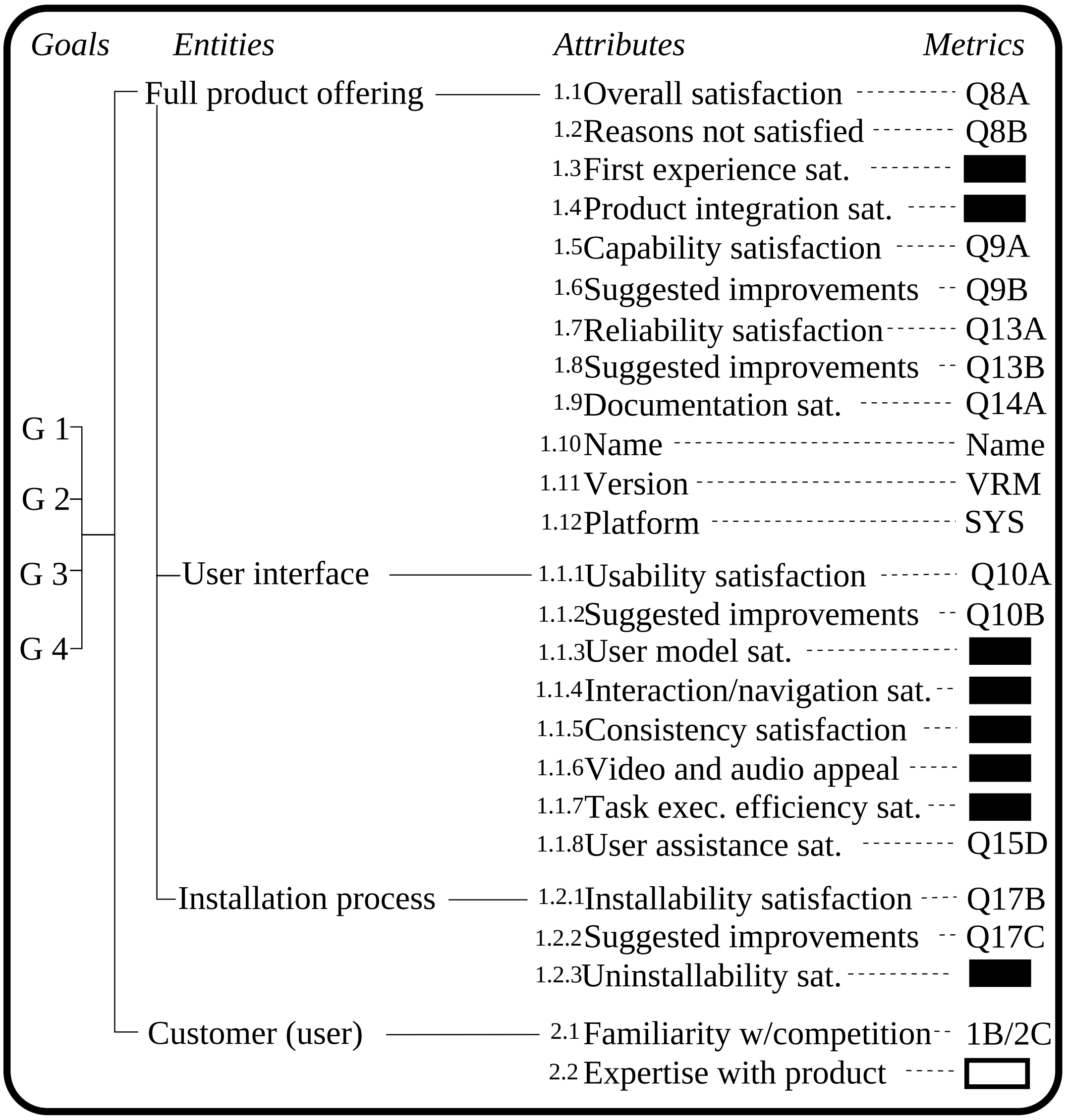

Figure 4.3: GQM Structure for the Usability Group

in surveys. Most of the missing metrics (1.1.3-1.1.7) make the bulk of a new attribute group – the usability attributes.

4.3 Bottom-up Analyses in the CUSTSAT MF

Our objective in applying the bottom-up analysis to the CUSTSAT MF is to extract new knowledge from the existing customer satisfaction data. More specifically, the data is explored to:

1. gain new business insights.

2. learn how the CUSTSAT information can be better collected and used.

Following the Section 3.3.1 guidelines, the following method was used to apply the AF Technique to the CUSTSAT data: (1) define an AF analysis using the generic relationship questions (GRQs); (2) collect and format the data for the analysis; (3) run the analysis; (4) review and organize diagrams; (5) interpret the resulting diagrams. The first and fifth steps do require the participation of an "expert." The CUSTSAT data manager at Toronto was responsible for defining the AF analyses and interpreting most of their results. The most interesting results were then shown to the data users during the data manager's periodical CUSTSAT data presentations.

This dissertation discusses AF analyses executed in 1995 and 1996. The 1995 analyses were planned based on some interesting results[2] obtained in 1994, when we first experimented with the AF tool. Although these "1995 analyses" were planned in the Fall of 1995, they were actually ran between December 1995 and February 1996 when I had already returned to the University of Maryland. During this period, I ran many analyses before the experts could review the diagrams. This proved to be a mistake.

In the data manager's first review of these diagrams (in June 1996), he commented that the diagrams were not showing clear results. We end up not extracting any major piece of knowledge from this large set of analyses. We concluded that the "1995 analyses" produced a large number of low level facts that did not lead to much knowledge discovery. This led us to adopt a more structured data mining approach for the "1996 analyses." We decided to first focus on the most important attributes. Those that influenced the laboratory's business decisions the most. We also decided to use data sets that involved whole product classes together. Classes involving all the products for a certain platform (e.g, mainframe, workstation, and PC products) or application type (e.g., compiler and database

[2]The data manager wanted to investigate what was the impact of the customer characteristics (factual information) on the CUPRIMDS (subjective) attributes.

Attribute	Meaning	Attr. Class
VENDOR	Vendor name	Vendor
PROD_TYPE	Product appl. type (DB, AD, or other)	ProdType
PLATFORM	Product Platform (PC, WK, MR, MF)	ProdType
Csat	Satisfaction with capability	SatA
Usat	Satisfaction with ease of use	SatA
Psat	Sat w/response time performance	SatA
Rsat	Sat with reliability	SatA
Dsat	Sat with documentation	SatA
Osat	Overall satisfaction with product	SatA
UPGRADE	Likelihood of upgrading	SatA
VENDORsat	Overall satisfaction w/vendor	SatA
RATING	Rating versus other products	SatA
REC_PRODUCT	Likelihood of recommending	SatA

Table 4.6: Attributes Used in the Vendor × ProdType × SatA Analysis

products) together. Our objective was to first explore the important data at a coarse granularity and later refine the analyses according to the obtained results. The following sections present the "1996 analyses." They summarize the interesting results extracted by the data manager during the 1996 diagram review interviews.

4.3.1 AF Analysis 1 – Satisfaction Attributes × Product Classes

The first 1996 analysis involved all SWS products and their competitors. The data manager set the following GRQ for this analysis: How do "Vendor" and "ProdType" affect "SatA ?" Our objective was to identify the satisfaction attributes (SatAs) in which IBM was better or worse than the competition. The ProdType attribute class is used to detail the satisfaction by platform and product application type[3]. The attributes used in the other attribute classes are listed in Table 4.6.

The VENDOR, PROD_TYPE, and PLATFORM attributes were especially derived for this first analysis. The VENDOR attribute has only two values – IBM (Lotus+IBM) and Competition (all others). It was derived from the product

[3]The product type was added after an initial analysis trial that did not produce many results. This initial trial also included the decision makers satisfaction attributes. We split this trial in the analyses describe here and in the next section.

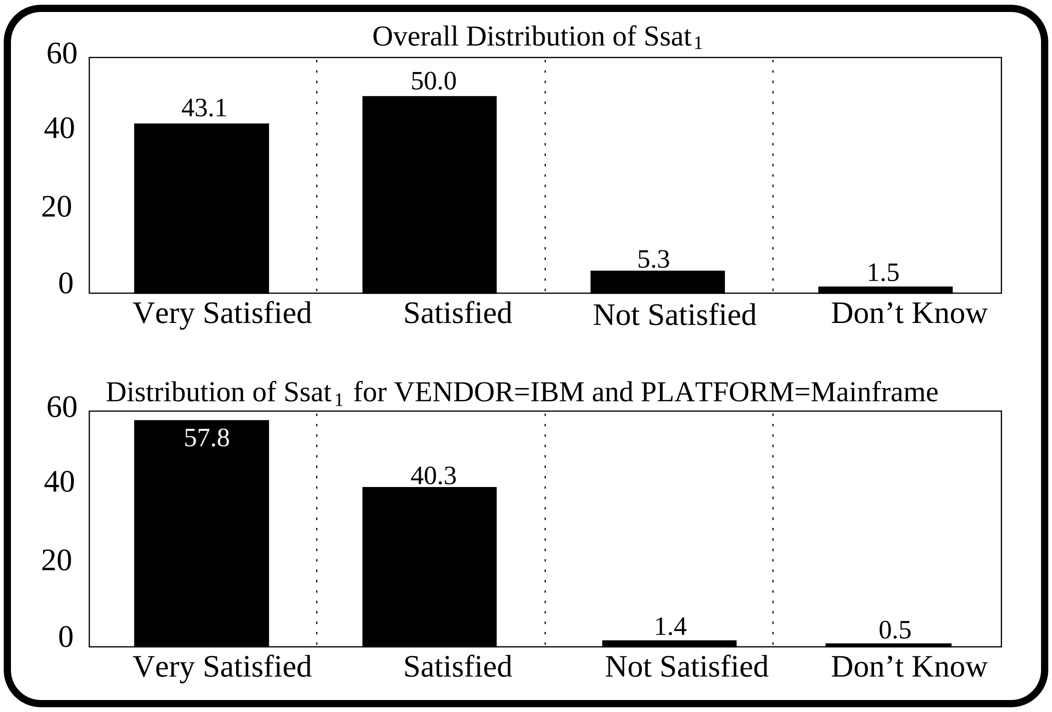

Figure 4.4: Most "Interesting" Diagram Produced in the First 1996 Analysis

vendor name. PROD_TYPE has three values – AD (application development – compiler), DB (database), and OTHER (all others). It was derived from the product name. PLATFORM has four values – PC (Personal Computer), WK (Workstation), MR (Mid Range), and MF (Mainframe). It was derived from the product's operating system name.

This analysis was run over the 1995 North America CUSTSAT data. A total of 8385 data points were used. Overall, the AF tool produced 19 diagrams for the Vendor $\times$ ProdType $\times$ SatA analysis. An auxiliary analysis – involving only ProdType $\times$ SatA – was also run. It produced 13 diagrams that helped us to interpret the original analysis. These diagrams were grouped using the algorithm showed in page 49. They were revised by the data manager in the end of June. The insights gained from them are listed in Appendix C.1.

The main result produced from this analysis is shown in Figure 4.4. In general, IBM had a significantly better performance than the competition in a certain attribute of its products. Let us call it $Ssat_1$. The PLATFORM attribute showed us that this advantage originated from the mainframe platform. In other words, IBM products were much better than the competition with respect to $Ssat_1$ in the mainframe platform and had similar scores in other platforms.

This result is interesting because $Ssat_1$ is a very important feature. However,

Attribute	Meaning	Attr. Class
VENDOR	Vendor name	Vendor
PROD_TYPE	Product appl. type (DB, AD, or other)	ProdType
PLATFORM	Product Platform (PC, WK, MR, MF)	ProdType
PRICEsat	Satisfaction with price	DMsatA
TERMSsat	Sat w/terms and conditions	DMsatA
COMP_PRICEsat	Sat w/competitive pricing	DMsatA
WILL_REPURCHASE	Would repurchase product	DMsatA

Table 4.7: Attributes Used in the Vendor × ProdType × DMsatA Analysis

the result was not novel. The data managers already knew about it. In this case, there was no knowledge discovered. Nonetheless, this was an illustrative experience for us. Two years ago, the data managers ran many statistical analysis to discover this very same information. The bottom-up method was able to find this interesting fact in a fast and cheap way.

4.3.2 AF Analysis 2 – Decision Makers Sat Attributes × Product Classes

The second 1996 analysis also involved all SWS products and their competitors. It was done together with the previous analysis and involved the decision makers satisfaction attributes (DMsatA). The decision makers are those customers that were involved is the decision to buy the surveyed product. The set up for this analysis is similar to the previous one. We used the the following GRQ: How do "Vendor" and "ProdType" affect "DMsatA ?" The attributes used in the DMsatA attribute class are listed in Table 4.7.

This analysis was run over the 1995 North America CUSTSAT data. A total of 3673 data points were used. Overall, the AF tool produced 9 diagrams for the Vendor × ProdType × DMsatA analysis. An auxiliary analysis – ProdType × DMsatA – produced 2 diagrams that helped us to interpret the original analysis. All these diagrams were grouped as before. They were revised by the data manager in the end of June. The insights gained from them are listed in Appendix C.2.

4.3.3 AF Analyses 3-6 – Local Support × MIAs

Next, a set of analyses was run to study the impact of the local support satisfaction (LSsat) attributes on the most important (MIA) satisfaction attributes. The analyses involved all North America customers that had used the local sup-

Attribute	Meaning	Attr. Class
LS_WHO	Who provided local support	LSwho
PROD_TYPE	Product appl. type (DB, AD, or other)	ProdType
PLATFORM	Product Platform (PC, WK, MR, MF)	ProdType
LS_SALESsat	Sat w/local sales support	LSsatA
LS_TECHsat	Sat w/local technical supp.	LSsatA
LS_EDUsat	Sat w/local education supp.	LSsatA
UPGRADE	Likelihood of upgrading	MIA
VENDORsat	Overall satisfaction w/vendor	MIA
Osat	Overall satisfaction with product	MIA
RATING	Rating versus other products	MIA
REC_PRODUCT	Likelihood of recommending the product	MIA
WILL_REPURCHASE	Would repurchase	MIA

Table 4.8: Attributes Used in the 1996 Local Support Analyses

port in the last 6 months. The attribute "LSwho" was used to determine who gave the local support: IBM, Competition, or Third Party. The attribute classes used in these analyses are shown in Table 4.8. The MIA class contains the most important attributes from the senior management point of view. Using the four attribute classes in Table 4.8, we ran the following analyses:

1. How does "LSwho" affect "MIA ?" [diagrams 1-6]
2. How does "ProdType" affect "LSsat ?" [diagrams 7-11]
3. How does "LSsat" affect "MIA ?" [diagrams 12-21]
4. How do "LSwho" and "LSsat" affect "MIA ?" [diagrams 22-55]

1191 data points – 591 for IBM and 600 for competition – were used in each analysis. With respect to the local support provider, the data is divided as 479 supported by IBM, 435 supported by competition, and 277 supported by third party. Overall, the AF tool produced 55 diagrams for these analyses. These diagrams were grouped by explained attribute by analysis. They were reviewed by the data manager in three distinct meetings. The insights gained from them are listed in Appendix C.3.

As seen in Appendix C.3, this analyses provided several interesting results. Here are some of the most intriguing ones:

- The competition may improve the ratings for some MIAs significantly if they improve their local support (Appendix C.3, Business Insights 15-18).

Attribute	Meaning	Attr. Class
Csat	Satisfaction with capability	Fsat
Usat	Satisfaction with ease of use	Fsat
Psat	Sat w/response time performance	Fsat
Rsat	Sat with reliability	Fsat
Dsat	Sat with documentation	Fsat
LS-Sales	Sat w/local sales support	Fsat
LS-Tech	Sat w/local technical supp.	Fsat
LS-Edu	Sat w/local education supp.	Fsat
UPGRADE	Likelihood of upgrading	MIA
VENDORsat	Overall satisfaction w/vendor	MIA
Osat	Overall satisfaction with product	MIA
RATING	Rating versus other products	MIA
RECOMMEND	Likelihood of recommending	MIA
REPURCHASE	Would repurchase product	MIA

Table 4.9: Attributes Used in the 1996 Fsats $\times$ MIAs Analysis

- A good technical support provided by a third party may have an enormous positive impact in one of the MIAs (Appendix C.3, Business Insight 12).
- In general, the LSsats showed very high associations with the MIAs.

The data manager was particularly intrigued by the last result. He always considered the CUPRIMDS attributes to be most important drivers of the MIAs. This result raised the question: are the LSsats as important as the CUPRIMDS with respect to the MIAs ? This question generated the following analysis.

4.3.4 AF Analysis 7 - CUPRD and Local Support $\times$ MIAs

The next analysis was designed to compare the impact of the CUPRIMDS and local support attributes on the MIAs. The CUPRIMDS and local support attributes were combined in one attribute class called Fsat. The Fsat and MIA classes are listed in Table 4.9. Isat, Msat, and Ssat were not included because only part of the customers answer these questions. Their null values would compromise the interestingness (attribute association) calculations.

The analysis – Fsat$\times$MIA – involved all North America customers that had used the local support in the last 6 months. Overall, we used 1191 data points – 591 for IBM and 600 for competition. The analysis was run with zero cutoff and asked for the first 40 of the 48 possible diagrams (8 Fsats $\times$ 6 MIAs). Diagrams were initially grouped by MIAs and interestingness level. However, we noticed that the Fsats had very different negative and positive impacts on the MIAs. For example, the first diagram produced was "$Csat \times MIA_1$." Its interestingness level was determined by the fact that for "$Csat$=very satisfied," "MIA_1=very satisfied" rose from 33.2 to 67.2%. So, what made this diagram interesting was Csat positive impact on MIA_1. On the other hand, "$Rsat \times MIA_1$" interestingness level was determined by the fact that for "$Rsat$=not satisfied," "MIA_1=not satisfied" rose from 8.6 to 42.6%. So, what made this diagram interesting was Rsat negative impact on MIA_1. For this reason, we decided to review the diagrams twice: (1) once with the diagrams ordered by Fsat positive impact on the MIAs; and (2) once with the diagrams ordered my Fsat negative impact on the MIAs.

In order to improve visualization, the facts expressed in the diagrams were summarized on Table 4.10. This table shows the positive and negative impacts of the Fsats on the MIAs. The positive impact was determined by the percentage of "very satisfied" (VS) answers for a MIA attribute given that the customers were "very satisfied" with a Fsat attribute. The negative impact was determined by the percentage of "not satisfied" (NS) answers for a MIA given that the customers were "not satisfied" with a Fsat attribute. The MIA's original VS and NS percentages are shown on the table's first column. The stars (*) mark the strongest impacts. The question marks (?) flag unreliable[4] impacts.

In order to preserve the IBM proprietary information, the MIAs are not explicitly identified in Table 4.10. The order of MIA attributes in Table 4.9 are not related to the order of MIA attributes in Table 4.10. The insights gained from Table 4.10 are listed in Appendix C.4. Here are some of the most intriguing ones:

- For some MIAs, LSsats are sometimes as important as the factors like product performance or reliability. For example, Table 4.10 shows that "local support sales" (a LSsat) has a higher positive impact than "reliability" with respect to MIA_3.

- The same attributes had different types of impacts in different MIAs. For example, "local support sales" was one of the attributes with the highest positive association with MIA_5, while it was one of the attributes with the lowest positive associations with MIA_1. This was a surprise because there used to be an implicit assumption that the Fsats were associated in more or less the same way with different MIAs.

- The same attributes may have quite different positive and negative impacts in the same MIAs. For example, "reliability" has a very high negative impact and a surprisingly low positive impact in MIA_4.

[4]This happens when the value distribution of a MIA varies significantly from when it is drawn from all possible values, to when it is drawn from the non-null values of a given Fsat.

MIA	Pos. Impact	New VS	Neg. Impact	New NS
VS=33.2%	Csat (*)	67.2%	Rsat (*)	42.6%
	Psat (*)	65.6%	Csat (*)	41.4%
MIA_1	Usat	61.6%	Psat	29.3%
	Dsat	59.0%	Usat	29.3%
	Rsat	57.6%	LS-Tech	20.8%
NS=8.6%	LS-Sales	51.9%	Dsat	20.2%
	LS-Edu	51.8%	LS-Sales	15.4%
	LS-Tech	50.3%	LS-Edu	14.3%
VS=18.5%	Dsat (?)	33.0%	Csat (*)	70.2%
	Csat	31.8%	Psat	59.3%
MIA_2	Usat	31.1%	Rsat	57.5%
	Psat	30.5%	Usat	53.3%
NS=26.8%	Rsat	27.4%	Dsat	45.6%
VS=53.9%	LS-Sales (*)	72.3%	Csat (*)	29.9%
	Psat	67.2%	Rsat (*)	27.6%
MIA_3	Csat	66.9%	Psat (*)	25.0%
	LS-Tech	65.9%	Usat	21.0%
NS=11.5%	Rsat	64.1%	LS-Sales	18.6%
	Usat	63.3%	LS-Tech	17.8%
	Csat (*)	62.5%	Rsat (*)	48.3%
VS=36.0%	Psat (*)	61.6%	Csat	40.3%
	Dsat (*)	61.1%	Usat	32.3%
MIA_4	Usat (*)	60.5%	Psat	28.6%
	LS-Tech	57.2%	LS-Tech	28.6%
NS=11.1%	LS-Sales	57.0%	Dsat	24.2%
	Rsat	54.7%	LS-Sales	23.7%
	LS-Edu	54.4%	LS-Edu	19.2%
VS=27.8%	LS-Sales (*)	42.1%	Rsat (*)	20.7%
	Psat	39.4%	Csat	16.1%
MIA_5	Usat	38.5%	Psat	12.9%
	Csat	38.2%	Usat	11.4%
NS=4.7%	Rsat	35.3%	LS-Sales	9.6%
	Csat (*)	72.1%	Csat (*)	44.9%
VS=48.7%	LS-Sales (*)	69.4%	Rsat	38.0%
	Psat (*)	68.9%	Usat	37.7%
MIA_6	Rsat	66.6%	Psat	35.0%
	Usat	66.4%	LS-Tech	28.6%
NS=13.3%	LS-Tech	65.6%	Dsat	27.8%
	LS-Edu	64.5%	LS-Sales	21.8%
	Dsat	61.1%	LS-Edu	16.8%

Table 4.10: Fsats × MIAs Results

Attribute	Meaning	Attr. Class
Csat	Satisfaction with capability	CUPRIMDS
Usat	Satisfaction with ease of use	CUPRIMDS
Psat	Sat w/response time performance	CUPRIMDS
Rsat	Sat with reliability	CUPRIMDS
Isat	Sat with installability	CUPRIMDS
Msat	Sat with maintainability	CUPRIMDS
Dsat	Sat with documentation	CUPRIMDS
Ssat	Sat with service support	CUPRIMDS
UPGRADE	Likelihood of upgrading	MIA
VENDORsat	Overall satisfaction w/vendor	MIA
Osat	Overall satisfaction with product	MIA
RATING	Rating versus other products	MIA
RECOMMEND	Likelihood of recommending	MIA
REPURCHASE	Would repurchase	MIA

Table 4.11: Attributes Used in the 1996 DB CUPRIMDS $\times$ MIAs Analysis

These facts led to more than new business insights. They showed that some assumptions about the data were incorrect or incomplete. They implied that some of the data analyses and models needed to be revised or refined.

These facts also led to the question: which is the most important MIA for the organization ? (Appendix C.4, MF Insight 6). This question cannot be answered with the data available in the MF, but it may be used to define new measurement goals. It is an AF result that can be mapped back to new GQM-based interviews.

4.3.5 AF Analysis 8 – CUPRIMDS $\times$ MIAs

The next analysis was designed to compare the impact of the CUPRIMDS on the MIAs: CUPRIMDS$\times$MIA. The CUPRIMDS and MIA attribute classes are listed in Table 4.11. The analysis involved all North America data points on "database" products. Overall, we used 682 data points – 499 for PC, 69 for workstation, and 114 for mainframe databases. The analysis was ran with zero cutoff and asked for the 48 possible diagrams (8 CUPRIMDS $\times$ 6 MIAs). Like the previous analysis, diagrams were grouped by positive and negative impacts on the MIAs.

In order to improve visualization, the facts expressed in the diagrams were summarized on Table 4.12. Like the previous one, this table shows positive and negative impacts of the CUPRIMDS on the MIAs. The stars (*) mark the strongest impacts. Questions used to measure Ssat, Isat, and Msat are only answered by customers that have used the service support, are installers, or are

MIA	Pos. Impact	VS	Neg. Impact	NS
	Psat (*)	59.9%	Csat (*)	35.2%
VS=25.4%	Usat	55.6%	Rsat (*)	34.8%
	Msat	54.9%	Ssat	21.6%
MIA_1	Csat	53.4%	Usat	21.4%
	Dsat	48.5%	Msat	20.4%
NS=8.1%	Isat	42.1%	Psat	19.3%
	Rsat	41.6%	Isat	16.7%
	Ssat	37.2%	Dsat	15.5%
	Msat (*) (?)	34.4%	Rsat (*)	74.0%
VS=16.0%	Psat (*)	32.3%	Csat (*)	66.2%
	Isat (?)	29.9%	Psat	51.6%
MIA_2	Usat	28.9%	Msat	48.1%
	Csat	28.4%	Usat	47.3%
NS=28.0%	Dsat	25.4%	Isat	41.0%
	Rsat	25.2%	Ssat	40.5%
	Ssat	21.8%	Dsat	39.1%
	Ssat (*) (?)	71.8%	Rsat (*)	34.8%
VS=52.5%	Msat (*) (?)	71.3%	Csat	23.9%
	Usat	65.9%	Psat	20.4%
MIA_3	Dsat	61.9%	Dsat	18.4%
	Csat	61.8%	Usat	18.3%
NS=14.2%	Isat	60.4%	Isat	17.9%
	Psat	59.9%	Msat (?)	14.8%
	Rsat	59.6%	Ssat (?)	10.8%

Table 4.12: $MIAs \times CUPRIMDS$ (Part a)

MIA	**Pos. Impact**	**VS**	**Neg. Impact**	**NS**
VS=32.0%	Msat (*)	59.0%	Rsat (*)	47.9%
	Usat (*)	57.0%	Msat	29.6%
	Psat (*)	56.9%	Csat	28.2%
MIA_4	Csat	53.9%	Ssat	21.6%
	Dsat	51.5%	Usat	19.8%
NS=10.7%	Rsat	50.6%	Isat	19.2%
	Isat	47.7%	Psat	17.2%
	Ssat	43.6%	Dsat	17.2%
VS=34.0%	Msat (*) (?)	58.2%	Csat (*)	22.5%
	Isat (?)	50.2%	Rsat (*)	21.7%
	Usat	48.9%	Psat	17.2%
MIA_5	Psat	48.5%	Msat	14.8%
	Ssat	46.1%	Usat	13.7%
NS=5.6%	Rsat	46.0%	Ssat	13.5%
	Csat	44.1%	Dsat	9.2%
	Dsat	39.5%	Isat	7.7%
VS=41.1%	Msat (*) (?)	63.1%	Csat (*)	46.5%
	Usat (*)	63.0%	Rsat (*)	43.5%
	Psat (*)	62.3%	Msat	38.9%
MIA_6	Csat (*)	61.8%	Usat	35.9%
	Dsat	58.2%	Psat	35.5%
NS=16.4%	Ssat	56.4%	Ssat	35.1%
	Rsat	55.0%	Isat	24.3%
	Isat	52.3%	Dsat	23.6%

Table 4.12: $MIAs \times CUPRIMDS$ (Part b)

maintainers, respectively. Their positive and negative impacts are computed using sub-sets of the total data set. They can be artificially inflated or deflated, if the value distribution for the MIA varies with those data sets. The questions marks (?) indicates where this happened.

In order to preserve the IBM proprietary information, the MIAs are not explicitly identified in Table 4.12. The insights gained from Table 4.12 are listed in Appendix C.5. Here are some of the most intriguing ones:

- For database products, performance and usability seems to have a very high positive impact in some MIAs.

- Reliability has low positive and very high negative impacts on all MIAs. The high negative impact was expected but the low positive impact was not.

- The very high maintainability impacts on the MIAs led to a very interesting insight about the MF itself. A CUSTSAT data manager of another IBM laboratory has raised the hypothesis that some customers may be misinterpreting "maintainability" as the ability to maintain the database instead of the ability to maintain the product.

If the hypothesis raised in the last insight is true, the survey would not be measuring what the data managers want the maintainability question to measure. Because of this insight, the maintainability question will be closely monitored on future surveys.

Chapter 5

Validation

Our work addresses three key issues: (1) better understanding the on-going measurement; (2) better structuring it; and (3) better exploring the data that the organization has already collected. It does not intend to be a comprehensive or definitive approach to improve measurement frameworks. As listed in Section 1.2.1, our work objectives are:

O1- discovering interesting data distributions and associations in the MF database

O2- visualizing data distributions and associations in the MF database

O3- assessing the importance of metrics for specific user groups and for the organization as a whole

O4- assessing the structure (i.e., measurement instrument, scale, and domain value) of metrics used in the MF

O5- assessing the appropriateness of the data collection process

O6- assessing the importance of data analyses for specific user groups and for the organization as a whole

O7- understanding and documenting the needs of users with respect to existing metrics, data analyses, and data presentations

O8- understanding and documenting the measurement goals of the MF data users

O9- identifying new applications and user groups for the data

O10- identifying the need for new metrics, data analyses, and data presentations

5.1 Validation Goals

We do not claim that our approach completely fulfills all the objectives listed above. The validation of our work aims to:

1. Determine if those objectives are really important for improving a measurement framework.
2. Evaluate the degree to which our approach fulfilled those objectives via the case study
3. Evaluate the cost at which our approach fulfilled those objectives in the case study

These issues can be expressed as the following GQM goals:

G1. Analyze the improvement objectives in order to evaluate them with respect to relevance from the data manager's point of view.

G2. Analyze the new approach in order to evaluate it with respect to effectiveness from the data manager's and data users' points of view.

G2.1. Analyze the existing improvement process in order to characterize it with respect to effectiveness from the data manager's and data users' points of view.

G2.2. Analyze the new approach in order to characterize it with respect to effectiveness from the data manager's and data users' points of view.

G3. Analyze our approach in order to evaluate it with respect to cost from the data manager's and data users' points of view.

5.2 Validation Process

A set of objective and subjective validations was performed to achieve the goals:

V1. In order to achieve the first validation goal (relevance of the objectives), the data manager was asked to subjectively judge how important each of the listed objectives is to improving the CUSTSAT measurement framework.

V2. In order to achieve the second validation goal (approach effectiveness), we:

V2.1 asked the data manager to: (1) subjectively judge the effectiveness of the phases that compose our approach in fulfilling the listed objectives, and (2) compare them with the current ad-hoc improvement process.

V2.2 compared the direct impact of the use of the approach on the CUSTSAT measurement framework with its ad-hoc improvement process.

V3. In order to achieve the third validation goal (the approach cost), we:

V3.1 asked the data manager to subjectively judge how cost effective the three steps of the approach were.

V3.2 measured how much effort was needed to apply the steps that compose our approach, and compared it with the effort to apply the ad-hoc improvement process.

V1 is referred to as the validation of the objectives relevance, V2.1 and V2.2 are referred to as the validation of the approach effectiveness, and V3.1 and V3.2 are referred to as validation of the approach cost effectiveness. V1, V2.1, and V3.1 are based on subjective evaluations. V2.2 and V3.2 are based on objective evaluations.

5.3 The Subjective Validation Questionnaire

The data for validations V1, V2.1, and V3.1 was collected jointly through one questionnaire submitted to the data manager. The aim of this questionnaire was: (1) to check how important the approach objectives are (V1); (2) to check how effective the ad hoc process is in fulfilling those objectives and how much each step of our approach has contributed towards fulfilling them (V2.1); and, (3) to check how cost effective our approach was (V3.1).

The questionnaire was divided by subject in six parts:

Part 1: Knowledge discovery and data visualization.

Part 2: Questions and questionnaire format evaluation.

Part 3: Assessment of the importance of questions, data analyses and data presentations.

Part 4: Understanding of user needs and goals.

Part 5: Identification of new user groups and definition of new questions, data analyses, and data presentations.

Part 6: Overall cost effectiveness evaluation.

The questionnaire used in the subjective validation is shown in Appendix D. The mapping between its questions and the approach objectives and validation goals is annotated in *italic* font in the questionnaire itself.

5.4 Importance of the Improvement Objectives

The validation questionnaire shown in Appendix D was used to evaluate the importance of the improvement objectives listed in the beginning of this chapter (V1). For that, the questionnaire has five point ordinal scale questions at the beginning of each section. These questions use the letters (A), (B), (C), (D), and (E) to quantify the improvement objectives. Option (A) meaning that the improvement objective has no importance at all. Option (E) meaning that the improvement objective has absolute importance.

Intelligent data exploration and knowledge discovery (O1) were considered of "some importance" (may be "of great importance") to the data manager business (C+). Visualization of data (O2) was considered of "great importance" (D) to the data manager business.

The ability to evaluate the metric's usefulness (O3) was considered of "absolute importance" (E) to the data manager business. The ability to evaluate the DA/P's usefulness (O6) was also considered of "absolute importance" (E). The ability to evaluate the structure of the metrics (O4) was considered somewhere between "of great importance" and "of absolute importance" (D+). The ability to evaluate the questionnaire structure (O5) was considered of "great importance" (D). The ability to identify new metrics (O10) was considered of "absolute importance" (E). The ability to identify new DA/Ps (O10) was considered of "great importance" (D).

The ability to understand user goals (O8) was considered of "great importance" (D). The ability to understand user needs (O7) was considered of "absolute importance" (E). Last but not least, the ability to identify new applications and user groups for the data (O9) was considered of "great importance" (D) by the data manager.

Figure 5.1 summarize the importance scores. It shows that, according to the data manager's subjective opinion, all the improvement objectives listed before are very relevant to the CUSTSAT MF.

5.5 Methods Effectiveness

The new approach was evaluated objectively by comparing the results obtained by its methods against the ad hoc improvements done to the MF during the

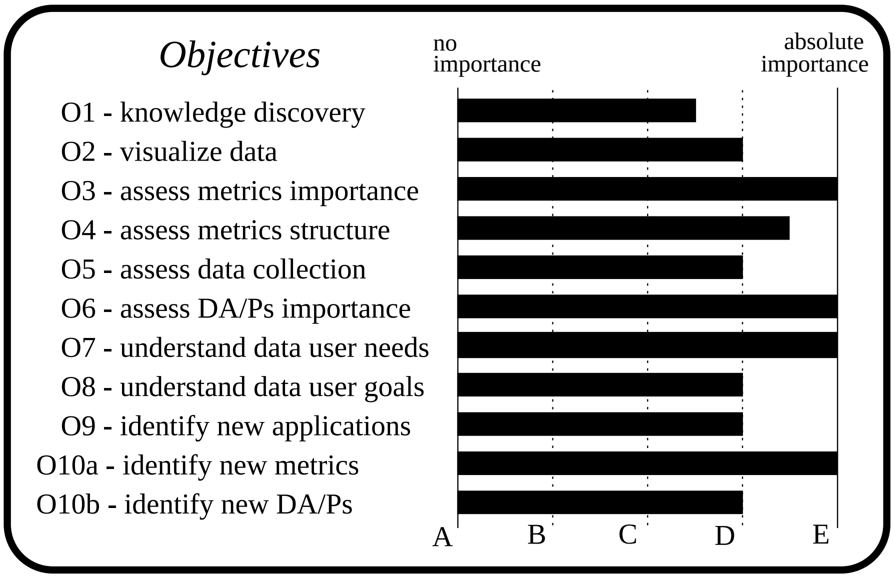

Figure 5.1: Subjective Rating of the Improvement Objectives

1995-97 period (V2.2). It was also evaluated subjectively by the answers given by the Toronto Lab data manager (V2.1) in the validation questionnaire.

5.5.1 Objective Evaluation

In order to validate V2, the impact of the new approach on the CUSTSAT survey questionnaire and on other parts of the measurement framework will be analyzed. The data for validation V2.2 was directly collected as follows:

- Impact on the CUSTSAT questionnaire
  - Survey questions (SQs) modified per user group. Relates to objective O4.
  - New SQs per user group. Relates to objectives O3 and O10.
  - Dropped SQs per user group. Relates to objectives O3 and O10.
  - Questions relocated inside the questionnaire. Relates to objective O5.
- Impacts on data usage
  - New data user groups. Relates to objective O9.
  - New data uses (DA/Ps). Relates to objectives O6 and O10.
  - Reviewed DA/Ps. Relates to objectives O6 and O10.

yr.	# of questions	**Method**	**Suggested**					**Implemented**		
			MQs	NQs	NQa	DQg	DQo	MQi	NQi	DQi
95	6	Ad hoc	NA	NA	NA	NA	NA	0	4+2	0
96	12	GQM	0	8	6	4	0	0	4	0
		Ad hoc	NA	NA	NA	NA	NA	0	3	0

Table 5.1: Results of the GQM Interviews with the Service Support Group

- Insights gained from data analyses. Relates to objectives O1 and O2.

These factors were evaluated for each of the three phases of the new approach, and compared to the ad-hoc MF modification process. Due to the small number of data points intrinsic to this type of case study, more attention is spent to identify the nature of the results obtained than to quantify them. The analysis done in this section is mostly qualitative.

Impact on the Questionnaire

Table 5.1 shows the comparison between the results of the GQM interviews with the service support (SS) group and the ad hoc process. The first column shows the year the method was applied. The second column shows the number of survey questions that were used by the group in that year's questionnaire. The third column shows the method name (GQM or ad hoc). The next five columns list the number of modifications suggested by the GQM method to the survey questionnaire. The following three columns show the modifications effectively implemented in the questionnaire.

At this point, it is important to remark that the modifications have to be approved by a committee before they can be implemented. This "CUSTSAT committee" is composed of CUSTSAT data managers from all SWS laboratories (Toronto, Raleigh, Santa Teresa, and Germany) and representatives from sales, headquarters, and manufacturing.

Table 5.1 compares modifications prompted by the GQM interview with one SS group against modifications prompted ad hoc by all SS groups of the SWS laboratories (Toronto, Raleigh, Santa Teresa, and Germany). Table 5.1 registers just the modifications actually made ad hoc to the CUSTSAT questionnaire. It was impossible to track down all the ad hoc modification suggestions, because there is no organized record of those suggestions inside the SWS laboratories. This information is marked as not available (NA) in Table 5.1.

Let us now look at Table 5.1 in more detail. Inside the "suggested" and "implemented" columns there are eight sub-columns:

MQ stands for modified questions (in the scale or range of value.)

MQs stands for question modification suggested.

MQi stands for question modification implemented.

NQ stands for new questions.

NQs stands for new question suggested.

NQa stands for new question suggested and adoptable.

NQi stands for new question implemented.

DQ stands for dropped questions.

DQg stands for questions found to be droppable for the interviewed group.

DQo stands for questions that are clearly droppable overall.

DQi stands for questions that were dropped.

The first line of Table 5.1 indicates that in the 1995 questionnaire there were six questions that were considered of interest to the service support group. Of those, none were modified or dropped. The numbers "4+2" in this line indicates that six new questions were implemented in the 1996 questionnaire and considered possibly useful to the SS group. Four of these were written specifically for the group (based on ad hoc requests). The other two were general questions thought to be useful to the SS group and later (in the GQM interview) considered extraneous by them.

The number of questions believed to be useful to the SS group in the following year (12) is equal to the summation of the number of questions in the first line: 6 existing plus 6 new questions.

The GQM method was applied to the SS group in 1996. Looking at the GQM structure for this group (Figure 4.1), one can identify 8 missing metrics. Six of these can be implemented and two are considered difficult to measure. This was represented by $NQ_s = 8$ and $NQ_a = 6$ in the table. One can also identify 4 generic questions that are extraneous from the SS point of view. None of these can (yet) be considered extraneous overall. This is represented by $DQ_g = 4$ and $DQ_o = 0$ in the table.

As the GQM structure for the service support group was available in 1996, it was used as one of the inputs for the 1996/1997 questionnaire modification meetings. The "implemented" column indicates that 4 of the 6 questions suggested by the GQM method were effectively implemented ($NQ_i = 4$). Two of the six attributes associated with black rectangles in Figure 4.1 were not effectively

yr.	# of questions	Method	Suggested					Implemented		
			MQs	NQs	NQa	DQg	DQo	MQi	NQi	DQi
95	18	Ad hoc	NA	NA	NA	NA	NA	1	1+1	1
96	19	Ad hoc	NA	NA	NA	NA	NA	1	2	1
97	20	GQM	1	5	4	8	2	NA	NA	NA
		Ad hoc	NA	NA	NA	NA	NA	NA	NA	NA

Table 5.2: Results of the GQM Interviews with the Documentation Group

adopted by the CUSTSAT committee. The attribute "most disliked attribute" was considered to be already covered by the attribute "suggested improvements" (Q18E). The attribute "satisfaction with commitment level" was considered difficult to measure and of particular interest to the service support group at the Toronto laboratory.

The third line of Table 5.1 shows that, besides the 4 metrics suggested by the GQM method, 3 other new metrics were added to the 1997 questionnaire because of ad hoc requests. Those metrics were added due to a request from a SS group from another laboratory. They are questions about the new Internet support activities (later adopted by all SS groups).

Table 5.2 shows the comparison between the results of the GQM interviews with the documentation group (ID) and the ad hoc process. Table 5.2 does not show the number of modifications effectively implemented in the questionnaire. This is because the interviews with this group were performed in 1997. The modifications suggested by these interviews will only be discussed by the CUSTSAT committee in the beginning of 1998.

When compared with the SS group's questions, fewer ad hoc modifications were done to the ID group's questions. There were two new questions implemented in 1995. One of them was later found to be extraneous for the group. As previously, this is represented by $NQ_i = 1 + 1$ in Table 5.2. Two other questions were implemented in 1996 based on an ad hoc request from the Toronto ID group. These metrics are the questions about tutorials shown in Figure 4.2.

The tutorial questions were found to be extraneous in the 1997 GQM interview. In fact, they were considered extraneous overall, and will probably be dropped in the 1998 questionnaire. Besides these two questions, six others were considered extraneous for the ID group only. This is represented by $DQ_g = 8$ and $DQ_o = 2$ in Table 5.2. As shown by the rectangles in Figure 4.2, the 1997 ID GQM interview also produced 5 new metrics. Four of these were considered "easy" to measure by the Toronto Lab data manager. This is represented by $NQ_s = 5$ and $NQ_a = 4$ in Table 5.2. The four "adoptable" questions will be

yr.	# of questions	Method	Suggested					Implemented		
			MQs	NQs	NQa	DQg	DQo	MQi	NQi	DQi
95	17	Ad hoc	NA	NA	NA	NA	NA	0	0	1
96	16	Ad hoc	NA	NA	NA	NA	NA	0	0	0
97	16	GQM	0	9	8	0	0	NA	NA	NA
		Ad hoc	NA	NA	NA	NA	NA	NA	NA	NA

Table 5.3: Results of the GQM Interviews with the Usability Group

submitted to CUSTSAT committee to be added to the 1998 questionnaire.

Table 5.3 shows the comparison between the results of the GQM interviews with the usability group (UI) and the ad hoc process. In 1995 and 1996, there was just one modification in the CUSTSAT questionnaire that affected metrics of interest to the UI group. The 1997 GQM interview has logged nine new metrics that are of interest to this group. Eight of them are considered easy to measure and will be submitted to the CUSTSAT committee in 1998. This is represented by $NQ_s = 9$ and $NQ_a = 8$ in Table 5.3. The eight "adoptable" questions will be submitted to CUSTSAT committee to be added to the 1998 questionnaire.

The insights gained with the AF analyses also suggested some modifications in the questionnaire. AF Analysis 5 made the data manager consider measuring the importance of the local support attributes (Appendix C.3, MF Insight 2). These two new questions were not adopted in the 1997 questionnaire. They will be discussed in 1998. Analysis 5 also raised the suggestion of moving the local support questions closer to the CUPRIMDS questions in the questionnaire (Appendix C.3, MF Insight 4). This modification was not done in 1997 but will also be discussed in 1998. Analysis 8 pointed to a possible problem of misinterpretation of the maintainability question for database products (Appendix C.5, MF Insight 4). This question will be monitored closely in the trial period (initial two weeks) of the 1998 survey.

Evaluation of the Impacts on Questionnaire

In order to evaluate the impacts of the new approach in the questionnaire, let us discuss the context and meaning of these impacts. The service support group made 4 ad hoc metric requests in 1995. This was the main reason we decided to interview them in the first place. However, the number of requests made by the SS group is not the rule but the exception. Except when a group is starting or stopping to use the CUSTSAT data, there are not many questions adopted in or dropped from the questionnaire for a given user group. The GQM interview with the Toronto SS group effectively produced 4 new metrics in 1996. It missed, however, the three questions related to the Internet service support that were later requested by a SS group from another laboratory.

The ID group is also a very active group in using the CUSTSAT data. However, their question set is more stable. There were few modifications on their question set in 1995 and 1996. In this scenario, the GQM interview contributed with suggestions to adopt 4 new questions and drop 2 existing questions from the 1997 questionnaire. The suggestion to drop 2 questions is of special interest because data users rarely request this type of thing in an ad hoc fashion. They ask for new metrics but usually do not communicate to the data manager that they do not need these metrics anymore. In this aspect, the GQM structures may help to keep the questionnaire from getting bigger than it needs to be. They enable the data managers to keep track of the data users present and past question needs.

The situation in the UI group is a bit different. They did not use the CUSTSAT data as frequently as the other groups. This is reflected by the number of ad hoc modification requests in 1995 and 1996; it was very low. The 1997 GQM interview coincided with a new corporate push for usability measurement and bridged the gap between user needs and measurement. This produced the suggestion of 8 new metrics to the 1998 questionnaire.

Considering these scenarios, the GQM interviews seemed very effective in proposing modifications to the questionnaire. Although the GQM interviews were done with only three groups, they produced an impact in the questionnaire comparable to the ad hoc requests from all similar groups inside the four SWS laboratories.

The AF analyses helped little to effectively modify the questionnaire. Its main contribution in this area was to help the data manager better understand the nature of local support attributes and to flag a possible problem with the maintainability question.

Impact on the Data Analysis

The GQM interviews produced some impacts on the regular data presentations. Both the UI and ID representatives asked for the inclusion of confidence intervals in the diagrams comparing the CUPRIMDS of IBM against the competition. The ID group also asked for new comparisons between IBM and the best product of the competition.

The AF-based method also suggested new data usages. The analysis of the CUPRIMDS against the most important attributes (MIAs) will be repeated for database and compiler product classes. These analyses will be presented to the several database and compiler user groups. They will include positive and negative impacts of the CUPRIMDS over the MIAs and will be repeated on a yearly basis.

Besides these new "regular" analyses, the AF-based method also suggested some exploratory data analysis. AF Analysis 7 originated from an insight gained during the AF Analysis 5 (Appendix C.3, Analyze Further 3). Similarly, AF Analysis 7 made the data manager consider doing a regression including local support and CUPRIMDS questions versus overall satisfaction (Appendix C.4, Analyze Further 4). And, AF Analysis 1 made the data manager consider a new analysis to find out which product is driving IBM scores for the $SatA_5$ attribute in the mid range platform (Appendix C.1, Analyze Further 1).

Modifications were also done ad hoc to the data usage during the 96/97 period. Due to a corporate push on the use of customer satisfaction data, the compiler groups are now having more frequent reviews of the CUSTSAT data. These reviews emulate the ones done by the database groups (the "CUPRIMDS/Osat monthly" and the "CUSTSAT annual" reviews described in Table 4.5).

The database groups' review of the customer comments on usability, documentation, and service support was also improved. All low satisfaction ratings and comments associated with the laboratory products are now copied to a particular database. Representatives of the ID and SS groups periodically use the database to call the customers and follow up on their comments. This is done because the comments obtained through the CUSTSAT survey are not specific enough to give appropriate feedback to these groups.

Evaluation of the Impact on Data Analysis

Let us qualitatively compare the modifications on data usage suggested by the ad-hoc process, the GQM interviews, and the AF analyses. All modifications that originated from the ad-hoc process were done to improve existing data usages, or to adopt data usages that emulated what was done by other user groups inside the laboratory. Although modifications in the DA/Ps are freely requested ad hoc, the GQM interviews did produce some new improvement suggestions. This indicates that this type of interview is an useful medium to improve existing DA/Ps.

The AF analyses were very effective in suggesting new and interesting DA/Ps for the CUSTSAT data. In this aspect, the AF-based method was clearly more effective than the ad hoc and GQM-based methods. The modifications suggested ad hoc and during the GQM interviews dealt with the improvement of existing data usage. In this sense, the AF-based method is clearly complementary to the other two. In the ad hoc and GQM-based approaches, the user focuses on improving the usage they are already making of the data. These approaches are driven by the immediate user needs. The AF-based method points to new possible data usages. It is driven by the new insights that are gained from exploring the data.

Insights Gained from Data Analyses

This section compares the type of insights gained from the regular data analyses and from AF data analyses. These comparisons are only qualitative because the regular data analysis and AF data analyses are different in nature. Regular data analyses are aimed at monitoring data considered important to the organization, in order to make business decisions based on changes in this data. The AF analyses were aimed at exploring the existing data in order to extract interesting facts from it. These "interesting" facts led to business insights, to new questions, or to insights about the MF itself.

In the CUSTSAT MF, regular analyses monitor key satisfaction areas. They examine IBM against the competition with respect to these satisfaction areas in order to determine if the gap between them is getting better or worse with time. New business insights are gained when the gaps between IBM and the competition change significantly. These insights are always important, but they are not frequent. For example, considering the regular analyses done with database products data in 1996, only two fundamental insights were gained from regular data analyses. It was found that the gap between IBM and competition had significant variations on two of the CUPRIMDS attributes during that year.

Instead of monitoring specific key areas, the AF analyses were aimed at finding new areas with interesting information. The AF analyses produced many and diverse insights on the data, but these insights were not always important. Appendix C lists the results of the AF analyses. Qualitatively the AF results were classified in three categories: require further analyses, produced MF insights, and produced business insights. Five results required or pointed to further data analyses. Eight results produced insights about the MF itself. Sixty one results produced business insights. The two fundamental types of insights gained with the AF analyses are discussed below:

- Insights about the MF itself: some of the modifications in the questionnaire and in data usage were suggested by AF analyses. They were listed and discussed in the two previous sub-sections. These insights were not numerous but they were important because they lead to better data utilization. In this aspect, the detection of new data uses is particularly important. Some of the most interesting results were the ones that pointed to new key areas that should be monitored in future data analyses. Examples of these areas are the positive and negative impacts of the CUPRIMDS attributes in the overall satisfaction, or the different impacts that the CUPRIMDS attributes have on different important attributes (MIAs).

- Insights about the business itself: several business insights were gained with the 1996 AF analyses. Examples are the discovery that good technical support given by a third party has a very strong positive impact in

MIA_4 (Appendix C.3, Business Insight 12), or that the MIA_6 scores for the competition may improve significantly if they improve local support (Appendix C.3, Business Insight 18). These business insights were much more common than the insights about the MF itself. However, not all of them are interesting. Furthermore, their importance varies with the user groups. Nonetheless, some of these facts did lead to important business insights any way one looks at them. An example of this type of important insight was the realization that for some of the MIAs the local sales support is as important as some of the key product attributes.

Evaluation of Insights

Our evaluation is that the AF and regular data analyses are complementary. Regular data analyses are aimed at monitoring key satisfaction areas. The insights gained with them are important but infrequent. AF data analyses are aimed at discovering new key satisfaction areas to be monitored. Their insights are much more frequent but only some of them are really important. Furthermore, the AF analyses did produce insights about the MF itself. This type of insight is very improbable in periodical regular analyses.

It is also important to note that AF insights also led to new measurement goals. Most of them were simple and could be directly mapped to a new data analysis. One of them – the data manager desire to identify which of the MIAs was the most important one (MF Insight 7 of AF Analysis 7) – cannot be achieved with the data available in the MF. If this goal is very important to the data manager, he might use the GQM-based method to identify what data should be collected to achieve it. This exemplifies how an insight gained from a bottom-up analysis can be fed back to a top-down data collection planning.

Other Results

A side result produced by the new approach that is worth mentioning is that the user groups and data uses documentation (produced during the characterization phase) were used as one of the inputs to the design of the CIS Web interface. This result exemplifies the usefulness of having explicit documentation about data uses and user groups in a MF.

5.5.2 Subjective Evaluation

The validation questionnaire shown in Appendix D was also used to evaluate the effectiveness of the new improvement approach (V2.1). The questions related to the effectiveness evaluation are marked as G2.1 (ad hoc process effectiveness) and G2.2 (the new approach effectiveness) in the questionnaire. These questions are quantitative and qualitative. The quantitative questions use a five point ordinal scale. The qualitative questions are open ended and ask for the data manager comments on the ratings he gave in the quantitative questions.

The aim of these questions is to qualitatively determine what our approach added to the ad hoc process with respect to the improvement objectives stated in beginning of this chapter. The main purpose of five point scale used in the quantitative questions was to make the data manager think about the issues we were discussing. They should not be taken as a quantitative stick of comparison between the ad hoc process and the new improvement approach. The questions were not formulated for this purpose and one interview is not enough to make this type of comparison.

In the results discussion below, the five point scale is represented by the letters (A), (B), (C), (D), and (E). Option (A) being the worst and option (E) being the best. The comments included in the text were made by the data manager during the interview.

Discovering Interesting Data Associations (O1)

According to the data manager the current ad hoc abilities to discover new interesting things in the data are "poor" (B). The only mechanisms available are the traditional statistical analysis packages.

The AF based method was rated as "very good" (E) for finding interesting data associations with respect to its impact on the current CUSTSAT MF activities. Its impact on major business decisions was considered only "good" (D), mainly because the CUSTSAT data is only one of the many inputs to those decisions.

The data manager commented that the AF method has broadened the scope of traditional analyses. It made it easier to look at a larger number of variables and analyze larger volumes of data.

Visualizing Data Distributions and Associations (O2)

The ad hoc process was considered "poor" (B). Usually the data is extracted and moved to spreadsheets from where data presentations are produced.

The AF-based method was considered "good" (D). According to the data manager, the diagram presentation in the AF tool can be improved. It would also be nice to be able to automatically produce result summaries like Tables 4.10 and 4.12.

Assessing the Importance of Metrics (O3)

The ad hoc process was considered "very poor" (A) both in assessing the metrics importance to specific user groups and to the organization as a whole. The questionnaire is available on the Intranet Web, but there is no mechanisms to force or prompt users to read through the questionnaire. The data presentations only discusses the data and not the questions used to collect them.

The MF characterization phase "helped significantly" (D) to picture which metrics were really important to which groups. However, it was useless (A) for understanding the importance of the metric in the organization as a whole.

The GQM-based method "helped a lot" (E) in understanding the metrics importance for the interviewed user groups. It "helped somewhat" (C) to understand that certain types of metrics were important to more than one group. For example, the metric "what did you like the most about the product ?" seems to be of general interest to several groups.

The AF-based method was "useless" (A) for understanding the metrics importance to specific user groups. For the organization as a whole, however, it "helped significantly" (D), especially if the results obtained during the 1994 (pilot) and 1995 analyses are included.

Assessing the Metrics Structure (O4)

The ad hoc process was considered "fair" (C) and needing of "much improvement." The ad hoc process involves the CUSTSAT committee and the survey vendor (who has a lot of experience with surveys). They look at comments or listen to interviews to check if the interviewees understand the survey questions. This process does not (but should) include data user representatives.

The MF characterization phase was considered "useless" (A) for assessing the metrics structure.

The GQM-based method was also considered "useless" (A) for assessing the metrics structure.

The AF-based method "helped a little" (e.g., the maintainability question for database products in Appendix C.5, MF Insight 7) to assess the metrics structure (B). However, the AF analyses focused on getting things out of the data as opposed to detecting problems with the metrics.

Assessing the Questionnaire Organization (O5)

The ad hoc process was considered "fair" (C), because it needs improvement. Every new questionnaire is reviewed by the CUSTSAT committee and the survey vendor. This year, correlations between all the questions are being run to assess the associations between them.

The MF characterization phase was considered "useless" (A) for assessing the questionnaire organization.

The GQM-based method was also considered "useless" (A) for assessing the questionnaire organization.

The AF-based method was also considered "useless" (A) for assessing the questionnaire organization. The data manager said he marked (A) instead of (B) because he would not consider the insight about modifying the location of the local support questions in 1997 (Appendix C.3, MF Insight 3).

Assessing the Importance of Data Analyses (O6)

The ad hoc process was considered good (D) for assessing the importance of data analyses to specific user groups and to the organization as a whole. In the current process, the data manager goes to the users and ask: "Is this analysis useful ?; Can we improve it in some way ?; What else would you like to see analyzed ?". Those inquiries are informal. The data manager also spends time discussing the optimal format of data presentations with senior representatives of the user groups. The data manager would like to see more formality without bureaucracy in this process.

The MF characterization phase was considered "useless" (A) for assessing the importance of data analyses. Here, it is important to mention that all insights gained during the user interviews were considered part of the GQM method by the data manager.

The GQM-based method "helped significantly" (D) for assessing the importance of data analysis to specific user groups and to the organization as a whole. The main reason for that is that the comments about the data presentations obtained during the user group interviews were considered quite helpful by the data manager.

The AF-based method was considered "useless" (A) for assessing the importance of data presentations. According to the data manager, the AF-based method did not affect existing data presentations, it helped to create new ones.

Understanding and Document Data User Needs (O7)

The ad hoc process was considered between "fair" to "good" (D–) for understanding and documenting data user needs. The data manager said that a quite formal process is followed to determine what products and competitors will be surveyed. The other aspects about the user needs (metrics and data presentations) are not covered so well. He said this happens because the product list is much more volatile than the questionnaire itself.

The MF characterization phase was considered between "of little help" and "of some help" (C–) for understanding and documenting data user needs. The descriptions of data uses and metric groups helped a little, especially by comparing data uses from the compiler and database groups.

The GQM-based method "helped significantly" (D) to understand and document data user needs. However, according to the data manager, the method focuses more on the metrics than on the data. For example, the list of products that should be surveyed was not discussed during the GQM interviews.

Understanding and Documenting Data User Goals (O8)

The ad hoc process was considered "very poor" (A) in understanding the data user goals. They do not have a process to do that. Because he is very experienced inside IBM, the data manager considers that he has a "fair" (C) understanding of the high level goals of the several data user groups. However, the MF has no explicit process to map these goals to the user needs (A).

The GQM-based method was considered "good" (D) for understanding and documenting user goals. According to the data manager the method "seems adequate," but he is not sure how good the method really is. He said the score could be (C), (D), or (E) depending on the goal accuracy. The mapping from goals to needs was considered "fair" (C). The GQM-based method maps goals to metrics but not to other user needs. For example, the goal purpose could be translated in more specific data collection and analysis needs.

Identifying New Applications for the Data (O9)

The ad hoc process was considered "good" (D) in identifying new applications and user groups for the data. According to the data manager, there exists mechanisms to do that: (1) periodic articles in internal news groups, magazines, and news letters mention the CUSTSAT framework and encourage new groups to use its data; (2) the main process document for software development inside IBM recommends the use of the CUPRIMDS measures; and (3) new areas and techniques of analyses suggested in current CUSTSAT conferences and journals are sometimes explored.

The MF characterization phase was considered "useless" (A) for identifying new applications and user groups for the data.

The AF-based method "helped significantly" (D) to find new applications for the data. According to the data manager, it did not help to find new user groups but it helped significantly to find new data and analyses for known user groups.

Identifying New Metrics and DA/Ps (O10)

The ad hoc process was considered "good" (D) to identify new metrics. Communication channels are kept open to the user groups. New suggestions are circulated inside the CUSTSAT committee. This process is effective in identifying and approving new questions.

The ad hoc process was also considered "good" (D) to identify new data analyses and presentations. Data presentations are considered an effective feedback channel to identify new data analyses for the user groups.

The AF-based method "helped somewhat" (C) to find new metrics to the framework. Some of the results suggested areas of discussion that may originate

Objective	Imp.	MC	GQM	AF	Ad hoc
O1	C+	A	A	D+	Weak mechanisms
O2	D	A	A	D	Weak mechanisms
O3(a) - user groups	E	D $(\Downarrow)$	E	A	Weak mechanisms
O3(b) - overall	E	A	C $(\Downarrow)$	D	Weak mechanisms
O4	D+	A	A	B	Some mechanisms
O5	D	A	A	A $(\Uparrow)$	Some mechanisms
O6(a) - user groups	E	A	D	A	Good mechanisms
O6(b) - overall	E	A	D	A	Good mechanisms
O7	D	C–	D	A	Some mechanisms
O8	E	A	D	A	Weak mechanisms
O9	D	A	A	D	Good mechanisms
O10(a) - new metrics	E	A	D $(\Uparrow)$	C $(\Downarrow)$	Good mechanisms
O10(b) - new DA/Ps	D	A	A $(\Uparrow)$	E	Good mechanisms

Table 5.4: Summary of the Subjective Evaluation

new metrics. The AF-based method "helped a lot" (E) to create new data analyses. The most interesting insights can be transformed in new data analyses. The notion of positive and negative impact of the CUPRIMDS on the MIAs was very significant to us (Appendix C.4, Business Insight 20, MF Insight 4).

The GQM-based method "helped significantly" (D) to identify new metrics. The user groups interviews were quite effective in producing new metrics. The GQM-based method was "useless" (A) to detect new data analyses and presentations.

5.5.3 A Final Analysis of Effectiveness

Table 5.4 summarizes the results from the subjective interview. Each row corresponds to one of the new approach improvement objectives. Objective O3 was split in: (a) assessing the importance of metrics for specific user groups, and (b) for the organization as a whole. Objective O6 was split in the same way, and Objective O10 was split in: (a) identifying new metrics, and (b) identifying new DA/Ps.

The first column (Imp.) shows the objective importance scores discussed in Section 5.4. The following three columns show the three phase of the improvement approach: the characterization phase (MC); the top-down analysis phase (GQM); and the bottom-up analysis phase (AF). The last column has the capabilities of the MF to achieve the listed objectives without the new approach. In order to indicate the different nature of the new approach and the ad hoc capabilities, the five point scores given by the data manager to the ad hoc process was transformed in a three point scale (weak, some, and good capabilities). The new approach scores correspond exactly to the subjective scores given by the data manager. They range from (A) to (E), (A) being very poor and (E) being very good.

The subjective scores given in the table can be compared to the objective evaluation of the results. Looking at Section 5.5.1, one can see that the objective evaluation pretty much supports the subjective scores. The points where the objective evaluation did not completely agree with the subjective scores were marked with arrows between parenthesis in Table 5.4.

The up arrow ($\Uparrow$) is used to indicate that the objective evaluation suggests a score higher than the one given by the data manager. Consider the case of GQM with respect to Objective O10(a) for example. Tables 5.1, 5.2, and 5.3 show that the GQM-based method was very good to identify new metrics when compared with the ad hoc process. This way the objective evaluation indicates that a score higher than (D) could be appropriate in this case.

Similarly, the down arrow ($\Downarrow$) indicates that the objective evaluation suggests a score lower than the one given by the data manager. Consider the case of AF with respect to Objective O10(a) for example. Except for MF Insight 2, there is very little evidence that the AF technique had an impact in identifying new metrics. A score lower than (C) seems more appropriate in this case.

Table 5.4 shows some important facts. The first one is that almost all the objectives listed were considered of great or of absolute importance to the MF. The second is that the MF is mature. It has capabilities in several of the areas that the new approach proposes to improve. In this aspect, a third fact should be highlighted. The capabilities that already existed in the MF are not the same as the ones provided by the new approach methods. The new approach complements or expands the MF capabilities even in areas where the MF already has good mechanisms to achieve the improvement objectives:

- For the Objective O10(a), identifying the need for new metrics, the MF uses the channels that are open between the data manager and the data users to successfully define new metrics. However, Tables 5.1, 5.2, and 5.3 show that the GQM-based method was very successful in producing new and locating extraneous metrics for the three user groups interviewed. This indicates that the GQM-based method was quite useful for detecting new and extraneous metrics in this mature MF.

- For Objective O10(b), identifying the need for new DA/Ps, and Objective O9, identifying new applications for the data, the MF uses the channels that are open between the data manager and the data user as mechanisms to successfully identify new applications for the existing data. Nonetheless, as discussed in Section 5.5.1, the DA/Ps proposed by the data users are aimed at emulating what is done by other user groups or at further exploring recognized key areas of the data. The AF-based method is aimed at discovering new areas to be explored and was quite successful in doing that. New DA/Ps and other applications for the data detected through the AF analyses are clearly a new asset to the organization.

- For Objective O6, assessing the importance of data analyses for specific user groups and the organization as a whole, the MF regular data presentations is a successful channel to assess the importance of DA/Ps. However, according to the data manager, the nature of the feedback gained with the GQM interviews is different than the one gained during regular presentations. More detailed and unbiased comments are gained during the GQM interviews.

- For Objective O7, understanding and documenting the user needs, the MF has a good process for determining and documenting the type of products that should be surveyed and the types of analyses that should be done with the data. However, they do not have good mechanisms to determine user needs with respect to the questions and questionnaire format. The GQM interviews capture exactly this information.

The fourth fact worthy of notice is that although all the improvement objectives were considered important, the MF has weak mechanisms to achieve some of them. This is true for: discovering interesting data distributions and associations (O1); visualizing the data (O2); assessing the importance of metrics for specific user groups and the organization as a whole (O3); and understanding and documenting the data user goals (O8). The new approach significantly helped to achieve those objectives. The data manager considered that the AF-based method helped significantly to achieve objectives O1, O2, and O3(b). He also considered that the GQM-based method helped significantly to achieve objectives O3(a) and O8.

The fifth fact worth noting is that the new approach "failed" to meaningfully achieve objectives O4 – assessing the structure of the metrics – and O5 – assessing the structure of the questionnaire. Although the objective results showed that the AF-based method has helped a little to assess the structure of the questionnaire, it is clear that, according to the data manager, the new approach does not help to find many problems with the structure of the questions and questionnaire.

The sixth fact worth mentioning is that the measurement characterization process did not help much to achieve the listed improvement objectives. This is not surprising as the main goal of the characterization phase is to document the MF key components in order to enable the bottom-up and top-down analysis phases.

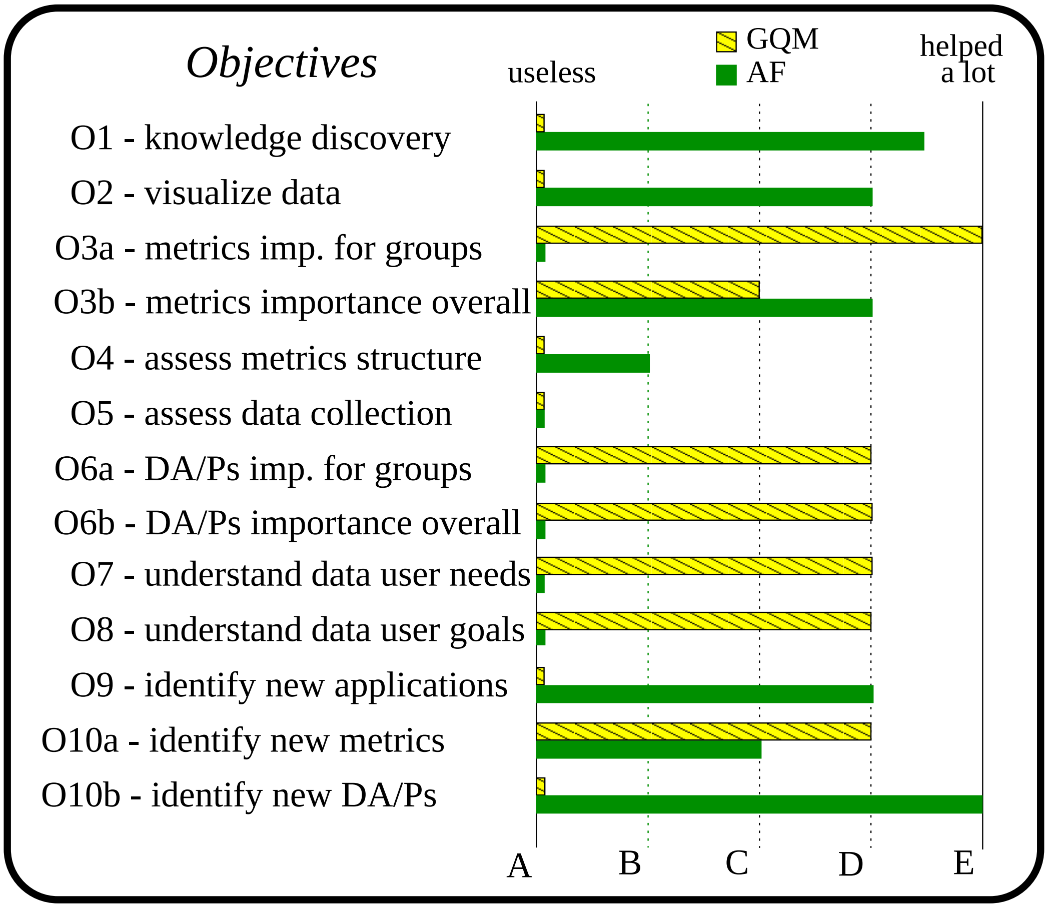

Figure 5.2: A Comparison Between the AF and GQM-based methods

Summary

The new approach complements or expand the MF capabilities even in areas where the MF already has good mechanisms to achieve the improvement objectives. The AF and GQM-based method helped significantly to achieve eight of the ten improvement objectives. More than that, they were quite complementary in achieving these objectives. The GQM-based method helped significantly to achieve objectives O3(a), O6(a), O6(b), O7, O8, and O10(a). The AF-based method helped significantly to achieve objectives O1, O2, O3(b), O9, and O10(b). This happened because the methods use complementary approaches to improve the measurement framework. The AF-based method works bottom-up. It uses the existing data as the driving force to improve the MF. The GQM-based method on the other hand works top-down. It uses the data user goals as the driving force to improve the MF. As shown in Figure 5.2, both methods were useful and contributed significantly to improve the measurement framework in several relevant and complementary aspects.

5.6 Cost and Cost Effectiveness

The cost analysis is also done objectively and subjectively. The objective analyses count the effort required to apply the methods and qualitatively compares it with the estimated effort to apply the ad hoc process. The subjective analysis asks the Toronto Lab data manager to evaluate how cost effective it was to apply the new improvement approach to the CUSTSAT MF.

5.6.1 Objective Cost Effectiveness Evaluation

In order to execute validation V3.2, we will check how much effort and time was spent to execute each of the steps of our method and compare it to the effort and time to execute the ad-hoc improvement process. The data for validation V3.2 will be directly collected as follows:

- Measurement framework characterization (MC) step

  C 1.1: Effort to document user groups and data uses.

  C 1.2: Effort to document metrics and available data.

- Bottom-up (AF) analyses

  C 2.1: Effort to plan and prepare data for each AF analysis.

  C 2.2: Effort to run the AF tool and organize the produced diagrams.

  C 2.3: Effort to review the produced AF diagrams.

- Top-down (GQM) analyses

  C 3.1: Effort to produce a tentative GQM structure for the user group interviews.

  C 3.2: Effort to review SQs, DA/Ps, and GQM structure with the data user group.

  C 3.3: Effort to organize and give the user feedback gained during the interviews back to the data managers.

- Ad-hoc process

  C 4.1: Effort spent handling and implementing questionnaire improvements requested by the data users and suggested by the data managers themselves.

	AN	DM	Activity	Code
1995	7.5	7.5	discussing data uses and user groups	C1.1
	6	—	documenting data uses and user groups	C1.1
	2	2	revising produced characterization	C1.1
	5	—	documenting metrics	C1.2
Total	20.5	9.5	= 30 person-hours	
1996	2.8	1.3	reviewing data uses and user groups	C1.1
	4	—	documenting data uses and user groups	C1.1
	0.7	0.7	discussing metric changes	C1.2
	2	—	documenting metrics	C1.2
Total	9.5	2	= 11.5 person-hours	

Table 5.5: Effort (in person-hours) Spent to Characterize the CUSTSAT MF

C 4.2: Effort spent in DA/Ps during the year.

Table 5.5 lists the costs associated with the characterization process in 1995 and 1996. Each row corresponds to a different activity. The columns list the effort of the MF analyst (myself) and the data manager, AN and DM respectively, in person-hours.

The characterization process in 1995 was labor intensive because it was the first time the MF was characterized using the process described in this dissertation. Almost half of the total effort was spend in interviews with the data manager (9.5$\times$2 person-hours). The process in 1996 was less labor intensive because the 1995 data served as the base for the new characterization. It basically represents the effort to capture and document the changes that occurred in the MF during the 95-96 period. Future updates of the characterization should require a similar amount of work.

Table 5.6 lists the costs associated with the 1996 AF analyses (AF A1 through AF A8). Each row corresponds to a different activity. The columns list the effort spent by the data analyst (myself), the data manager, and others (AN, DM, and OT respectively) in person-hours. The "others" category include members of the CUSTSAT committee and data users that were present in very short presentations of the AF A7 and A8's summary of results. The AF analyses were also labor intensive, especially to the data analyst. This effort can be reduced significantly with some improvements in the data mining tool. The effort to review and organize diagrams can be reduced by improving the diagrams preview, browse, and print facilities. The procedure to organize diagrams by negative and positive impacts can be automated. The effort to extract and format data was greatly reduced after the first analysis. This happened because there was a one

	AN	DM	OT	Activity	Code
AF	0.5	0.5	—	planning analysis	C2.1
A1	4.5	—	—	extracting and formatting data	C2.1
	3.5	—	—	running tool and organizing diagrams	C2.2
	3	3	—	reviewing the diagrams	C2.3
Total	11.5	3.5	—	= 15 person-hours	
AF	0.2	0.2	—	planning analysis	C2.1
A2	0.5	—	—	extracting and formatting data	C2.1
	0.5	—	—	running tool and organizing diagrams	C2.2
	1.5	1.5	—	reviewing the diagrams	C2.3
Total	2.7	1.7	—	= 4.4 person-hours	
AF	0.7	0.7	—	planning analysis	C2.1
A3-6	0.5	—	—	extracting and formatting data	C2.1
	3.5	—	—	running tool and organizing diagrams	C2.2
	6.5	6.5	—	reviewing the diagrams	C2.3
Total	11.2	7.2	—	= 18.4 person-hours	
AF	0.1	0.1	—	planning analysis	C2.1
A7	0.2	—	—	extracting and formatting data	C2.1
	3	—	—	running tool and organizing diagrams	C2.2
	3	—	—	organizing data by positive and neg. impacts	C2.2
	2.2	2.2	3.5	reviewing the diagrams	C2.3
Total	8.5	2.3	3.5	= 14.3 person-hours	
AF	0.1	—	—	planning analysis	C2.1
A8	0.2	—	—	extracting and formatting data	C2.1
	3	—	—	running tool and organizing diagrams	C2.2
	2	—	—	organizing data by positive and neg. impacts	C2.2
	2	—	—	analyzing inflated scores	C2.2
	1.5	0.5	4	reviewing the results	C2.3
Total	8.8	0.5	4	= 13.3 person-hours	
Grand total for the eight 1996 AF analyses = 65.4 person-hours					

Table 5.6: Effort (in person-hours) Spent to Run AF Analyses

	AN	DM	DU	Activity	Code
SS	2	0.5	0.2	arranging the interview	C3.1
	1	1	—	discussing the SS process	C3.1
	4	—	—	deriving a tentative GQM structure	C3.1
	6	—	—	designing the interview script	C3.1
	0.3	0.3	—	revising the interview script	C3.1
	1.5	—	1.5	interviewing data users	C3.2
	1.3	0.3	—	passing the interview results back to the DM	C3.3
Total	16.1	2.1	1.7	= 19.9 person-hours	
ID	2	0.5	0.2	arranging the interview	C3.1
	3	—	—	deriving a tentative GQM structure	C3.1
	4	—	—	designing the interview script	C3.1
	0.5	0.5	—	revising the interview script	C3.1
	1.5	—	2.5	interviewing data users	C3.2
	2.5	0.5	—	passing the interview results back to the DM	C3.3
Total	13.5	1.5	2.7	= 17.7 person-hours	
UI	2	0.7	0.4	arranging the interview	C3.1
	3	—	—	deriving a tentative GQM structure	C3.1
	2	—	—	designing the interview script	C3.1
	0.3	0.3	—	revising the interview script	C3.1
	3.0	—	4.0	interviewing data users	C3.2
	4.0	—	—	mapping the beta survey questions to the CUSTSAT survey questions	3
	3.3	0.3	—	passing the interview results back to the DM	C3.3
Total	17.6	1.3	4.4	= 23.3 person-hours	

Table 5.7: Effort (in person-hours) Spent to Produce GQM Structures

time effort to create a program to clean and format the data extracted from the CUSTSAT database.

Table 5.7 lists the costs associated with the GQM-based method. Each row corresponds to a different activity. The columns list the effort of the MF analyst (myself), the data manager, and the interviewed data users (AN, DM, DU respectively) in person-hours. The time required from the data managers and data users was very low. This is important because senior representatives of the data user groups are costly resources. In spite of this fact, the GQM interviews took a sizeable effort to prepare. However, almost half of the data analyst total effort was spent to design the interview script and to prepare a tentative GQM structure for the group. Both costs will be drastically reduced in future interviews with the same groups. In this scenario, interviewing data users in a periodical base (e.g., annually) could cut the cost of the listed GQM interviews by half.

For all laboratories involved in the survey, the yearly ad-hoc effort to modify and update the questionnaire (C4.1) was estimated to be around 120 person-hours during the three first months of the year. This effort is composed of: (1) teleconferences to discuss and approve modifications; and (2) monitoring the survey closely during its trial period. The customer satisfaction committee also meets two to four times a year to discuss the survey and the data. As these meetings involve seven people, this effort adds up to a number between 80 to 200 person-hours. This effort does not include the time spent to get feedback from users. The users give the data managers feedback on the questionnaire throughout the year, but this effort could not be estimated.

Inside the Toronto Lab, there are monthly data presentations to the compiler and database groups (C4.2). From the data manager point of view, these presentations takes four to eight person-hours to prepare and two person-hours to present. This adds up to number between 12 to 20 person-hours per month, or 96 to 160 person-hours per year (the CUSTSAT data is only collected during 8 months of the year). The effort of users to review the data is at least 320 person-hours (2 hours $\times$ 10 people $\times$ 2 main groups $\times$ 8 times per year). This number does not include the data users' effort to review the data by directly using the CIS Web Interface.

In total, an effort that ranges from 500 to more than 800 people-hours is spent annually on the ad hoc processes. The total effort spent to apply the new approach was around 170 person-hours. However, this effort does include startup and learning costs of the new methods. In this scenario, the new improvement approach can be considered worthwhile. Especially if one considers that the new approach has capabilities that were considered important and complementary to the capabilities that the MF already had.

5.6.2 Subjective Cost Effectiveness Evaluation

The cost effectiveness of each phase of the new approach was subjectively evaluated by the Toronto Lab data manager (V3.1). The cost effectiveness questions are located in the final section (Part 6) of the questionnaire shown in Appendix D. There were three questions for each phase of the new approach. There is one quantitative question using a five point scale (from A to E) and two open ended questions asking about the main benefits and drawbacks of the new approach methods.

The first question was about the MF characterization process. The data manager said that it was of "of modest value" (C). He does not know if the cost of applying it outweighs its benefits. According to him, the characterization "gave me the opportunity to stop the day to day business and look at the whole thing from a higher level." The process helped the data manager to picture where the MF is today, as opposed to when he consciously thought about it in the past.

The main drawback was that a lot of his effort was spent in the characterization phase.

As shown in Table 5.5, the characterization was one of the most costly phases of the new improvement approach from the data manager point of view. Table 5.4 shows that it had little direct impact on the improvement objectives. This justifies the data manager evaluation of the characterization phase. However, the data manager did not consider the fact that the MF characterization phase is a precondition to apply the bottom-up and top-down methods, especially if the person doing the analyses is not familiar with the MF.

The GQM-based method was considered "of considerable value" (E). Its main benefit was the user feedback that the data manager got from the GQM interviews. The proposed metrics and comments about the DA/Ps were considered good. The data manager also added that people are more willing to criticize the MF when they are talking to an independent party. The feedback he gained through the GQM interviews was not biased by his own opinions about the MF. The main drawback was that all this information was not directly obtained by him. He is concerned that important pieces of information might have been missed during the interviews. He also asserted that the GQM structures show the data user needs in terms of metrics and not in terms of data analyses and data sets to be collected.

The AF-based method was also considered "of considerable value" (E). The main benefit was that new insights were gained in the MF data in several different areas. Large amounts of data and variables could be analyzed quickly. According to the data manager, the method was able to come up with things that he would never be able to come up with on his own. There were two main drawbacks. The first was that the tool could use some improvement and be extended to produce summaries of results. The second was that the obtained results usually need to be further explored statistically to prove their significance. According to the data manager, the reactions of other members of the CUSTSAT committee to the AF results were also very positive.

Chapter 6

Conclusions

This dissertation argues that to understand and improve existing measurement frameworks is an important problem in software engineering. It introduces an approach to tackle this problem on two key fronts: (1) understanding and improving on-going measurement; (2) better exploring the data that the organization has already collected. A MF characterization process and a GQM-based method was used to tackle the first problem. A data mining (AF-based) method was used to tackle the second problem.

The GQM-based method is founded on the principles of goal-oriented measurement - more specifically on the GQM Paradigm. It is aimed at applying the principles of goal-oriented measurement in an environment that is already functional, instead the more common view of defining a measurement process from scratch based on the measurement goals of data users. It aims at assessing if the user goals can be fulfilled by the data that is already being collected.

The AF-based method uses a data mining to approach the problem from a different angle. Instead of improving the current measurement process, it improves data usage. It does that by discovering new interesting information in the existing data.

The new approach was tested in a case study performed in a real and mature measurement framework. We used it to improve the customer satisfaction measurement framework (CUSTSAT MF) at the IBM Toronto Laboratory (Toronto Lab).

6.1 Main Results and Work Contributions

The case study tested the three parts of the new approach with respect to ten MF improvement objectives. Those objectives are listed in Page 2. All of them were considered important or very important by the Toronto Lab data manager. The case study showed that the new approach was effective in achieving eight of these ten objectives.

The case study also showed the CUSTSAT MF is mature and has diverse mechanisms to achieve some of the improvement objectives. Nonetheless, even in this scenario, the AF and GQM-based methods contributed by complementing and expanding the capabilities that already existed to improve the MF. This indicates that the new approach acted in areas that were important but were being ignored in this and possibly in others MFs.

The characterization process tackled the problem of understanding how people are using the data in a measurement framework. The data manager considered that this process did not produce many improvements in the MF. Nonetheless, we believe characterization has a key role in the new improvement approach. It is a pre-requisite for applying the GQM and AF-based methods. When a MF is not well known or documented, characterization seems to be a fundamental first step in any effort to improve it. Although characterization is an important step in improving existing and legacy MFs, we do not know of another work that discusses this problem.

The GQM Paradigm has been used by several software engineering organizations. However, it has been used to plan and implement measurement from scratch. One of the main contributions of this dissertation is to show how the GQM Paradigm can be applied when the measurement framework is already operational. Our contributions focus on detecting missing and extraneous metrics.

The AF-based method describes how the AF technique can be applied in a measurement framework. Our contributions here are: (1) in the area of data mining, the dissertation proposes procedures to better explore data using the AF technique; (2) in the area of software engineering, the dissertation shows that this type of data exploration can produce important business insights and contribute to better understanding the data, metrics, and measurement models used in software organizations.

The GQM and AF-based approaches are complementary. The case study showed that the top-down and bottom-up analyses mutually complement each other with respect to the listed improvement objectives (see Figure 5.2). The GQM-based method improved the MF top-down driven by the data user goals. The AF-based method improved the MF bottom-up by the better exploring the existing data.

We believe that AF and GQM can also work in synergy. The GQM structures can be used to choose and organize data for AF analyses. The measurement goals can be mapped to generic relationship questions and used to define AF analyses. The AF results can be fed back to measurement goals and used to revise existing GQM structures. This should be done whenever AF analyses raise new questions that require further data analyses. If the data needed to run those analyses are not being collected inside the MF, the raised questions should be mapped to new measurement goals.

The new approach was designed to be non-intrusive to the MF management.

	Contributions
Software Engineering	the formalization of some important GQM concepts, such as the semantic of the goal facets and the question templates.
	the instantiation of the GQM Paradigm to improve existing measurement frameworks.
	the design of the case study used to evaluate the methods.
	the process defined to characterize existing MFs.
Data Mining	the association new methods with the AF Technique, in particular: (1) the use of generic relationship questions to define attribute classes to reduce the space searched by the AF Technique; (2) the use of attribute ordering to improve visualization of cause-effect relations in the AF diagrams; and, (3) the use attribute classes to define an algorithm to organize the AF diagrams.
	the use of generic relationship questions to create an interface between GQM (a measurement planning paradigm) and AF (a data mining technique).

Table 6.1: Main Contributions of This Work

Its main objective is NOT to implement modifications to a MF, but rather to point to where it can be improved. It is also important to point out that the new approach is not a methodology for defining new metrics or measurement (predictive) models. It is rather a methodology for understanding the data, the metrics, and how they are fulfilling the needs of data users.

Table 6.1 summarizes the contributions of this work to the software engineering and data mining fields.

6.2 Work Limitations

The new approach is not by any means a complete or definitive approach to improve a measurement framework. As seen in Chapter 5, it does have important capabilities that are complementary to the ones that already existed in the CUSTSAT MF. However, the new approach only contributes towards solving some of the problems that are associated with a MF.

The GQM-based method applies the GQM Paradigm incrementally – one point of view at a time. This approach has a limitation in detecting extraneous metrics. We can only detect metrics that are extraneous from a certain point of view. We have to interview all the user groups that are related to a certain metric before we can conclude that this metric is extraneous to the MF as a whole.

The GQM-based method focus mainly on detecting extraneous and missing metrics. It does not use the GQM goals to determine the data to be collected (e.g., it does not help to define the surveyed population and sample size). It also does not map goals to data analyses and decision making models.

The AF-based method cannot substitute for statistical data analysis techniques. It complements them. The bottom-up method gives us the ability to find interesting facts that might otherwise remain hidden in the data. It is geared towards "discovering" information. One should use statistics to further analyze the facts discovered using this method. A good discussion on how statistics should be combined with data mining is found in [56].

The AF tool used in the case study was designed to mine nominal and ordinal data. The tool has limitations in exploring interval and ratio data. Numeric-valued attributes have to be mapped into discrete ranges of values before they can be used in an AF-analysis. Consider the metric "lines of code" (LOC) for example. Its numeric values have to be mapped to a discrete and finite set of values if one wants to use them in AF analyses. Suppose that the values "small, medium, large, and very large" are considered adequate to quantify software size. In this case, one could define the metric "discrete software size" using the following mapping: (1) the discrete software size is "small" if $LOC < 5000$; (2) it is "medium" if $5000 \leq LOC < 100,000$; (3) it is "large" if $100,000 \leq LOC < 500,000$; and (4) it is "very large" if $LOC \geq 500,000$.

The case study worked with a very small number of data points, our conclusions could not be expressed quantitatively. This dissertation can only claim that the new approach brought new and important capabilities to a mature MF. Similar case studies with a larger number of data points must be run if one wants to study the new approach capabilities quantitatively.

6.3 Lessons Learned

During the data uses and user groups characterization interviews, it was easier for the data manager to describe the data uses first. It is easier to identify user groups from the list of data uses than to identify data uses from the list of user groups. We also noted that the amount and type of data usage varies even between user groups with similar goals. This creates the opportunity to identify data usage experiences that can be transfered between groups.

The GQM interviews should be done with senior representatives (e.g., managers) of the data user groups. Only senior representatives can state measurement goals and effectively identify relevant attributes. Users that work closely with the data but are not at the management level cannot do that. The senior representatives of the user groups are busy people and sometimes will recommend that MF analysts interview this type of user instead of him/her. This should be avoided.

AF analyses should explore the most important and generic data first. The data exploration can be refined later based on the insights gained in the earlier analyses. It is important to explore the data at a coarse granularity before exploring it at finer granularity levels. This allows the data analysts to understanding the data behavior at a higher (more generic) level before exploring it at lower (more detailed) levels.

6.4 Future Work

One natural extension of the characterization phase is to try to automate the data usage procedures described for each user group, like it was done in the CIS Web interface. An interesting extension would be to define a basic language to describe data uses in a way that they could be easily automated.

The GQM-based method can be expanded to produce mappings from goals to whole measurement and data use plans. This raises two important research questions: (1) how can GQM goals be used to define data analyses and decision making models ?; and (2) how can these models be integrated to software development and maintenance process models.

The AF-based method can be expanded with other types of visual and forensic data mining techniques. Besides trying to discover interesting data associations, one can try to detect interesting data sequences or clusters, or simply to create efficient and simple mechanisms to summarize and visualize the data.

One natural extension to the current AF-based method is to use assumptions to help detect interesting data associations. An assumption is a statement believed to be true about the relationship between attributes of interest [30]. Assumptions can be used as one of the inputs to the functions that calculates the interestingness of data associations. The more an association deviates from an assumption the more interesting it is.

We intend to further explore the synergy between AF and GQM, and to define an integrated method of reengineering measurement frameworks. We want to couple AF and GQM more tightly. Our idea is to better formalize the use of GQM to structure existing measurement frameworks, and combine it with different types of data mining approaches.

Appendix A

Data Collection Forms

A.1 MF Characterization Forms

Attribute Group

Filled by: Date:

Group Description:

Attribute	Description	Metric

Data Use

Filled by: Source: Date:

Name:

Description:

Frequency of Use:

Metrics Involved:

How Data Sets Are Selected:

Participant User Groups (*):

Comments

(*) Using the following classification, indicate between parenthesis how important is this data for each group:

1. Very important — they rely on it heavily to take business decisions.
2. Important — they rely on it substantially to take business decisions.
3. Relevant — sometimes they rely on it to take business decisions.
4. Not much relevant — they don't use the data regularly in their business
5. No relevance — they never use the data in their business

Data User Group

Filled by: Source: Date:

Group:

Role Description:

How important is the CUSTSAT data for this group ?

1. Very important — they rely on it heavily to take business decisions.
2. Important — they rely on it substantially to take business decisions.
3. Relevant — sometimes they rely on it to take business decisions.
4. Not much relevant — they don't use the data regularly in their business
5. No relevance — they never use the data in their business

How long have they been using the data ?

Metrics of Interest:

Purposes of Using Them (Data Uses):

Suggest Other Possible Uses:

Comments:

A.2 AF Analyses Forms

AF Data Set:

Filled by: Source: Date:

Description / Selection Criteria (include number of data points):

Originally Extracted to:

Metrics:

Data Year: **Geographic Region:**

Time to Extract: **File Name:**

Trials for:

Filled by: Source: Date:

Trial number:

Time to organize attributes:
Tool running time:
Time to inspect and organize diagrams:

Attributes/Grouping/Ordering:

AF Cut-off: **# Asked/Produced:**
Folder Name:

Trial number:

Time to organize attributes:
Tool running time:
Time to inspect and organize diagrams:

Attributes/Grouping/Ordering:

AF Cut-off: **# Asked/Produced:**
Folder Name:

List of Interesting Facts

Filled by: Source: Date:

Refer to Analysis:

Type: **A)** Business Insight **B)** Problem Report **C)** Further Analysis Required

Importance: **A)** Very Important **B)** Important **C)** Helpful **D)** Irrelevant

Diagrams Involved:

Facts Shown:

Possible explanations:

A.3 Effort Sheet

Time Sheet

Refers to:

Activity	Description	Participants	Date	Effort

Appendix B

Script Used During the Documentation Group Interviews

The following handout was used to interview representatives of the documentation group. In this handout, "ID" stands for Information Development (documentation group). Other acronyms used in it are part of the organization lingo.

B.1 Data Use and Its Importance

First I want to document how the ID group uses the customer satisfaction information, and how important this data is for you.

B.1.1 Regular data presentation

There is a regular presentation of the customer satisfaction data on the DB Products. During these meetings the following information is usually presented:

1. Comparison between the lab products' CUPRIMDS/O and the competition.
2. Comparisons between the lab products' NSI and the competition over time (YTD and YE).
3. Evaluation of the CUPRIMDS satisfaction vs importance.

Data use

Which of these analyses are useful to the ID group ? How do you use them ?

B.1.2 Direct Contact with Customer

I understand that you sometimes call dissatisfied customers or review their comments stored in the CIS database. With what frequency is this done ?

Data Use

What prompt you to call dissatisfied customers ? How useful is the information they provide you ? How do you use it ?

B.1.3 Other Data Uses

Besides the data uses listed in this section do you use the CIS data in any other way ? If yes, in what way ?

Are there any other application for the customer satisfaction data that you would like to pursue in the future ?

B.1.4 Intranet Access to CIS

Is the ID group using (or planning to use) the CIS intranet interface at *http://xxxxxxx.xxxxxx.xxx.xxx/cis2/* ?

If you are not, skip to Section B.1.5. If you are, continue.

Data Use

How are you using (or planning to use) the CIS intranet interface ?

B.1.5 Overall Importance of the CUSTSAT Data

Overall what is the importance of the customer satisfaction data to your group ?

1. Very important - we rely on it heavily to make business decisions.
2. Important - we rely on it substantially to make business decisions.
3. Relevant - sometimes we rely on it to make decisions.
4. Not much relevant - we don't use the data regularly to make decisions.

B.2 Needed Data

Now, I would like to identify what other useful information could be gathered to the group by the CUSTSAT survey. In order to do that, we will follow a top-down process. I will start with your goals in using the CUSTSAT data and finish with a list of attributes that we could measure for you. I understand that your generic goal in using the CUSTSAT data is to "improve the documentation with respect to customer satisfaction". Is that correct ?

The diagram I am showing you now is the result of the top-down process applied to the previous top-level goal.

Considering how you use the customer satisfaction data to improve your business, could you break the top level goal into more concrete sub-goals ?

B.2.1 Entities

Based on my understanding of the ID process, I have identified the following entities/objects as relevant to your top level goal:

- Documentation provided by vendor
  - Printed Manuals
  - On-line help screens
  - Soft-copy books
  - Tutorials
  - Others (e.g. technical newsletters)
- Documentation provided by others
- Translation process
- Customer contact
- Customer organization
- Product

Thinking of DB's products, services, people, processes, and activities that can be characterized or evaluated by the customers. Are there any other entities/objects that are relevant to your goals ?

B.2.2 Attributes

The diagram I gave you has attributes related to the top level goal. Some of them are already surveyed by the CUSTSAT questionnaire, others were added by me.

Three types of attribute were added: (1) attributes to evaluate the relative importance between the documentation deliverables; (2) attributes to recognize which other types of documentation are being used by the customers; and (3) attributes to evaluate the deliverables accuracy, completeness, and usability.

- Documentation provided by vendor

  1. Overall satisfaction with documentation (14a)
  2. Documentation importance (19e)
  3. Types of deliverables used (15a)
  4. Most important deliverable (??)

  - Printed Manuals
    1. Satisfaction with printed manuals (15b)
    2. 2 or 3 suggested improvements (15c)
    3. Satisfaction with the manuals accuracy (??)
    4. Satisfaction with the manuals completeness (??)
    5. Satisfaction with the manuals usability (??)
    6. Most liked aspect of the manuals (??)
    7. Importance of having printed manuals (??)
  - On-line help screens
    1. Satisfaction with on-line help screens (15d)
    2. 2 or 3 suggested improvements (15e)
    3. Satisfaction with the help accuracy (??)
    4. Satisfaction with the on-line help completeness (??)
    5. Satisfaction with the screens usability (??)
    6. Most liked aspect of the on-line help screens (??)
    7. Importance of having on-line help screens (??)
  - Soft-copy books
    1. Satisfaction with soft-copy books (15f)
    2. 2 or 3 suggested improvements (15g)
    3. Satisfaction with the books accuracy (??)
    4. Satisfaction with the books completeness (??)
    5. Satisfaction with the books usability (??)
    6. Most liked aspect of the soft-copy books (??)
    7. Importance of having soft-copy books (??)

- Tutorials
  1. Satisfaction with tutorials (15h)
  2. 2 or 3 suggested improvements (15i)
  3. Satisfaction with the tutorials accuracy (??)
  4. Satisfaction with the tutorials completeness (??)
  5. Satisfaction with the tutorials usability (??)
  6. Most liked aspect of the tutorials (??)
  7. Importance of having tutorials (??)
- Others (e.g. technical newsletters)
  1. Name (??)

- Documentation provided by others
  1. Name (??)

- Translation process
  1. Satisfaction with translations (14b)
  2. 2 or 3 suggested improvements (14c)
  3. Translation importance (20m)

- Customer contact
  1. Job responsibilities (45c)

- Customer organization
  1. Primary business (45a)
  2. Type of activities (45b)

- Product
  1. Name (P_Name)
  2. Version (P_VRM)

Please rate these attributes as: (1) very relevant; (2) relevant; or (3) not relevant to you.

Now, considering each of the entities/objects listed in the previous section, list other attributes that are relevant to you AND can be evaluated by a survey of customers (please, clearly explain the attributes meaning and indicate those that are very relevant to you).

B.2.3 Metrics

The 1997 CUSTSAT questionnaire has the following questions regarding the ID deliverables: Q14a-14c, Q15a-15i, Q19e, and Q20m.

Do you have any comments on the wording, ordering, or positioning of these questions in the questionnaire ?

I am giving you a complete list of the questions used in the 1997 customer satisfaction questionnaire. They are organized by subject. Please revise it, if you have time. Let us know, if there is any new data that interest your group.

Appendix C

Insights Gained from AF Diagrams

This appendix lists the insights gained during the 1996 AF Analyses. The attributes groups used in these analyses are explained in Section 4.3. In order to preserve IBM proprietary information, the explained attributes are represented by acronyms. The acromyns are composed by the attribute class name and an ordinal number. These numbers are not related to the order in which the attributes are presented in Section 4.3. We also use the phrase [IBM proprietary information removed] to indicate that a piece of text was cut to protect IBM proprietary information.

C.1 Results of the AF Analysis 1

This analyses is described in Section 4.3.1. It involved three attribute classes: "Vendor × ProdType × SatA". These three classes are described in Table 4.6.

This analysis produced 19 diagrams that were organized in 8 diagram groups based on the diagram's explained attribute ($SatA_N$).

Diagram Group 1: Diagrams 1 and 2 - $SatA_1$

Diagram 1 shows that IBM's scores for $SatA_1$ are much better than the competition in the mainframe platform.

Diagram 2 shows that IBM's scores $SatA_1$ are significantly better than the competition overall (very sat=47.8% against 37.5%).

Insights:

[Known fact] IBM's scores for $SatA_1$ are better than the competition overall, but this advantage is coming from the mainframe platform. The data manager already knew about this fact, but he was impressed by how easy it was to catch it with the AF Tool.

Diagram Group 2: Diagrams 3 and 19 - $SatA_2$

Diagram 3 shows that IBM compiler products' scores for $SatA_2$ are significantly better than the overall average.

Diagram 19 shows that in general IBM is a bit better that the competition with respect to attribute $SatA_2$.

Insights:

[Known fact] IBM's scores for $SatA_2$ are better than the competition overall, but this advantage is coming mainly from the application development products.

Diagram Group 3: Diagram 4 and 16 - $SatA_3$

Diagram 4 shows that IBM compiler products' scores for $SatA_3$ are significantly better than the overall average.

Diagram 16 shows that IBM mainframe products' scores for $SatA_3$ are significantly better than the overall average.

Insights:

[Known fact] IBM's scores for $SatA_3$ are especially strong in its application development and mainframe products. The data manager said this was expected.

Diagram Group 4: Diagram 5 and 18 - $SatA_4$

Diagram 5 shows that IBM PC products' scores for $SatA_4$ are significantly higher than the overall average.

Diagram 18 shows that in general the competition has lower scores for $SatA_4$.

Insights:

[Business Insight 1] IBM's scores for $SatA_4$ are better than the competition overall, but this advantage is coming mainly from the PC platform.

Diagram Group 5: Diagrams 6, 7, 13, 14, and 15 - $SatA_5$

Diagram 15 shows that in general the IBM products have better $SatA_5$ scores.

Diagram 6 shows that IBM 's $SatA_5$ scores are significantly worse than the average for the "others" product class.

Diagram 7 shows that IBM's $SatA_5$ scores are significantly better than the average for the application development product class.

Diagram 13 shows that IBM's $SatA_5$ scores are significantly worse than the average for the mid range platform.

Diagram 14 shows that IBM's $SatA_5$ scores are better than the average for the PC platform.

Insights:

[**Business Insight 2**] The IBM's $SatA_5$ scores are better than the competition but there is a problem with products in the mid range platform and in the "others" products class.

[**Analyze Further 1**] The data manager believes that some products may be driving these low scores. He was especially curious if the product "XXXX" was driving the low $SatA_5$ scores in the mid range platforms.

Diagram Group 6: Diagrams 8 and 11 - $SatA_6$

Diagram 8 shows that customer's $SatA_6$ scores given to IBM's compiler products are higher than others products in general.

Diagram 11 shows that customer's $SatA_6$ scores given to IBM's PC products are somewhat higher than other products in general.

Insights:

[**Known Fact**] IBM's $SatA_6$ is especially strong in its application development and PC products. The PC scenario is somewhat surprising, but the difference is not very strong. Auxiliary diagram 11 shows that the PC products in general have a slightly higher $SatA_6$ score.

Diagram Group 7: Diagrams 9 and 10 - $SatA_7$

Diagram 9 shows that customer's $SatA_7$ scores IBM's PC products are somewhat lower than other products in general.

Diagram 10 shows that customer's $SatA_7$ scores for IBM's mainframe products are somewhat higher than other products in general.

Insights:

[**Inconclusive Fact**] Looking at the auxiliary diagram 7, we concluded that diagram 10 is not significant because in general mainframe customers are more satisfied with respect to $SatA_7$ (nonetheless, IBM has a slight $SatA_7$ advantage in this platform).

[**Business Insight 3**] IBM's $SatA_7$ disadvantage in the PC platform is more significant.

Diagram Group 8: Diagrams 12 - $SatA_8$

Diagram 2 shows that IBM's $SatA_8$ scores is better than the competition overall.

Insights:

[**Known Fact**] The advantage of 5% that shift from not satisfied to very satisfied with the vendor is significant but appears to be coming from the mainframe platform and/or application development products.

C.2 Results of the AF Analysis 2

This analyses is described in Section 4.3.2. It involved three attribute classes: "Vendor × ProdType × DMsatA" Analysis. These three attribute classes are described in Table 4.7.

This analysis produced 9 diagrams that were organized in 3 diagram groups based on the diagram's explained attribute ($DMsatA_N$).

Diagram group 1: Diagrams 1 and 7 - $DMsatA_1$

Diagram 1 shows that IBM's application development products have a better $DMsatA_1$ rating than the overall average rating.

Diagram 7 shows that the competition's $DMsatA_1$ rating is worse than IBM's rating.

Insights:

[**Known Fact**] IBM's $DMsatA_1$ is better overall, but is especially better for compiler products.

Diagram group 2: Diagrams 2 and 3 - $DMsatA_2$

Diagram 2 shows that the competition's $DMsatA_2$ rating for the PC platform is significantly better than the overall average.

Diagram 3 shows that IBM's $DMsatA_2$ rating for the PC platform is slightly better than the overall average.

Insights:

[**Business Insight 4**] As shown by the auxiliary diagram 1, the products running in PC platforms have better $DMsatA_2$ ratings in general. We concluded that IBM has a worse $DMsatA_2$ rating than the competition in the PC platform by comparing diagrams 2 and 3.

Diagram group 3: Diagrams 4, 5, 6, and 9 - $DMsatA_3$

Diagram 4 shows the competition's $DMsatA_3$ ratings for PC products is higher than the overall average.

Diagram 5 shows the competition's $DMsatA_3$ ratings for WK products is lower than the overall average.

Diagram 6 shows IBM's $DMsatA_3$ ratings for PC products is slightly higher than the overall average.

Diagram 9 shows the competition's $DMsatA_3$ for DB products is lower than the overall average.

Insights:

[**Known Fact**] The auxiliary diagram 2 shows that the $DMsatA_3$ ratings changes mainly with the platform (PC, WK, MR, and MF). Most of those variation is explained by the platform rather than the vendor.

[**Known Fact**] In general, the products running in PC platforms have better $DMsatA_3$ ratings. We concluded that IBM has worse $DMsatA_3$ ratings than competition for the PC platform by comparing diagrams 4 and 6.

[**Business Insight 5**] Diagrams 5 and 9 seems to indicate that IBM does better than the competition in DB applications and especially in WK platforms.

C.3 Results of the AF Analyses 3-6

As explained in Section 4.3.3, the local support analyses involved four different analyses that produced 55 diagrams:

1. AF Analysis 3: $LSwho \times MIA$ [diagrams 1-6]
2. AF Analysis 4:$ProdType \times LSsat$ [diagrams 7-11]
3. AF Analysis 5:$LSsat \times MIA$ [diagrams 12-21]
4. AF Analysis 6:$LSwho \times LSsat \times MIA$ [diagrams 22-55]

The attribute classes used in these analyses are explained in Table 4.8. The produced diagrams were grouped by analysis and explained attributes. In order to save space below, the diagram groups and the insights gained with Analyses 3, 4, and 5 are grouped together. The diagrams produced by Analyses 6 were organized in several groups according to their explained attribute.

Diagram group 1: Diagrams 1 thru 6 - $LSwho \times MIA$

Diagram 1 shows that the MIA_1 is the highest when IBM is the local support (LS) provider and the lowest when a third party is the LS provider. The same trend is seen in MIA_2, and MIA_3 (diagrams 2 and 6).

Diagram 3 shows that the MIA_4 is the highest when IBM is the LS provider and the lowest when the competition is the LS provider. The same trend is seen in MIA_5 and MIA_6 (diagrams 4 and 5).

Insights:

[Known Fact] The facts in diagrams 3, 4, and 5 are expected because IBM products rate better overall in the MIAs. As the third party providers give support to IBM and competition products their score should fall in between the other two.

[Business Insight 6] The diagrams 1, 2, and 6 are surprising. They are showing that third party local support has a (small) negative impact on MIA_1, MIA_2, and MIA_3 scores. The third party score is particularly low on MIA_1. This shows thaat [IBM proprietary information removed] when a third party is the support provider.

Diagram group 2: Diagrams 7 thru 11 - $ProdType \times LSsat$

Diagram 7 shows that $LSsat_1$ is the highest on MR platform products. It decreases for the MF, PC, and WK platforms respectively. Diagrams 9 shows that $LSsat_1$ is the lowest for DB products and the highest for "Other" products.

Diagram 9 shows the "Other" product class have the best $LSsat_1$ scores and DB products the worse. Diagram 8 is not conclusive.

Diagram 10 and 11 show that $LSsat_2$ and $LSsat_1$ scores are the highest for DB products in the PC platform.

Insights:

The data managers doesn't have definitive explanations for the results.

[Business Insight 7; Analyze Further 2] Diagrams 7 is intriguing because the WK platforms have a particularly low $LSsat_1$ rating.

[Inconclusive Fact] Diagrams 9, 10, and 11 are not conclusive.

Diagram group 3: Diagrams 12 thru 21 - $LSsat \times MIA$

Sub-group 1: MIA_4

Diagrams 12, 16, and 17 show that good local technical and sales support have similar positive impacts on MIA_4. These impacts are bigger than the local education impact.

Sub-group 2: MIA_3

Diagrams 13, 15, and 19 show that good local technical, education, and sales support have similar positive impacts on MIA_3.

Sub-group 3: MIA_6

Diagrams 14, 18, and 20 show that good local sales has the biggest positive impact on MIA_6.

Sub-group 4: MIA_1

Diagram 21 shows that local sales "probably" has the highest impact in MIA_1.

Insights:

[Business Insight 8] It seems that local sales has strong impact on MIA_1 and MIA_6. This was surprising.

[Analyze Further 3] The data manager commented that local support in general had very strong impacts on the MIAs. He suggested that we analyzed them against the CUPRIMDS factors (this originated the analyses in Section 4.3.4).

[MF Insight 1] We discussed the possibility of adding importance questions regarding sales and education support to the questionnaire.

[MF Insight 2] We also discussed how these results supported the decision of combining the local technical support questions with the vendor technical support question (S_SAT) in the 1996 questionnaire.

[MF Insight 3] We also discussed that this might suggest that the other local support questions should be moved closer to the technical service support questions.

Diagram group 4: Diagrams 22-25, 32-33, and 34-36 - $LSwho \times LSsat \times MIA_3$

Diagrams 22 thru 25 (together with 6 and 23) show that very high satisfaction with local education has the highest positive impact on MIA_3 when IBM is the education provider.

Diagrams 32 and 33 (together with 6 and 19) show that very high satisfaction with local sales support has the highest positive impact on MIA_3 when IBM the provider.

Diagrams 34 thru 36 (together with 6 and 13) show that a very high satisfaction with third party local technical support may have positive or negative impacts in MIA_3. Very good technical support by third party providers produces more very satisfied and more dissatisfied scores with respect to MIA_3.

Insights:

[Business Insight 9] Diagrams 22 thru 25 seems to show that good local education is a bit more important to MIA_3 when IBM is the education provider.

[Business Insight 10] Diagrams 32 and 33 seems to show that good local sales is a bit more important to MIA_3 when IBM is the provider.

[Business Insight 11] Diagrams 34 thru 36 may indicate that good third party technical support produces better MIA_3 results in general. However, in some situations, only good vendor support can improve MIA_3.

Diagram group 5: Diagrams 26-29, 39-42, and 50-52 - $LSwho \times LSsat \times MIA_4$

Diagrams 26 thru 29 (together with 3 and 12) show that very good technical support provided by a third party has a very strong positive impact on MIA_4

Diagrams 39 thru 42 (together with 3 and 17) show that very good local sales provided by a third party may have positive or negative impacts in MIA_4. Very good sales support by third party providers produces more very satisfied and more dissatisfied scores with respect to MIA_4.

Diagrams 50 thru 52 (together with 3 and 16) show that very good education support by IBM has the strongest positive impact on MIA_4.

Insights:

[Business Insight 12] Diagrams 26 thru 29 show a very different trend between third party and other providers. It seems that when a third party gives good technical support the satisfaction with the MIA_4 sky rockets. A possible explanation is that the customer associates good operation record with the product and vendor but the problems with the support provider.

[Business Insight 13] Diagrams 34 thru 36 may indicate that good third party sales support produces better MIA_4 results in general. However, in some situations, only good vendor support can improve MIA_4.

[Inconclusive Fact] Diagrams 50 thru 52 are showing the same behavior as diagram 3. IBM rates are better overall, competition come last, and third party comes in between. The same trend holds for MIA_4 distributions when the customer is very satisfied with the local education.

Diagram group 6: Diagrams 30 and 46 - $LSwho \times LSsat \times MIA_1$

Diagram 30 (together with 1 and 21) shows that very good IBM sales support has a very strong impact on MIA_1.

Diagram 46 (together with 1) shows that very good IBM technical support has a very strong impact on MIA_1.

Insights:

[Business Insight 14] Diagram 30 and 46 shows that good sales and technical support by IBM have good impact on MIA_1. However, this impact is also present in less extent for the other providers. Note that according to diagram 1, MIA_1 scores are higher in general when IBM is the local support provider.

Diagram group 7: Diagrams 37-38, 43-45, and 53-54 - $LSwho \times LSsat \times MIA_6$

Diagrams 37 and 38 (together with 5 and 18) shows that MIA_6 is bit higher when the competition provides sales support and the customers are very satisfied with it.

Diagrams 43 thru 45 (together with 5 and 14) shows that MIA_6 is significantly higher when the competition provides tech support and the customers are very satisfied with it.

Diagrams 53 and 54 (together with 5 and 18) shows that MIA_6 is similar when IBM or the competition provides sales support and the customers are very satisfied with it.

Insights:

[**Business Insight 15**] Diagrams 37 and 38 are interesting because they contradict diagram 5. This may indicate that good sales support has a greater impact on the competition with respect to MIA_6.

[**Business Insight 16**] Diagrams 43 thru 45 are very interesting because they strongly contradict diagram 5. This may indicate that good tech support has a greater impact on the competition with respect to MIA_6.

[**Business Insight 17**] Diagrams 53 and 54 are interesting because they contradict diagram 5. This may indicate that good education support has a greater impact on the competition with respect to MIA_6.

[**Business Insight 18**] The diagrams 37-38, 43-45, 53-54, and 5 hint that MIA_6 ratings for the competition may improve significantly if they improve their local support.

C.4 Results of the AF Analysis 7

This analysis is described in Section 4.3.4. It involved two attribute classes: "Fsat $\times$ MIAs". These two classes are described in Table 4.9.

This analysis produced 40 diagrams that were organized in two sets of six diagram groups based on the diagram's explained attribute (MIA_N). The first set of diagrams was ordered by the Fsats positive impacts on the MIAs. The second set was ordered by the Fsats negative impacts on the MIAs.

Diagram group 1: MIA_1

Positive impact:

Table 4.10 shows that Csat and Psat are the attributes with the highest positive impact (PI) on MIA_1. Usat, Dsat, and Rsat (in this order) have medium PI. And, LS-Sales, LS-Edu, and LS-Tech have the lowest PI.

Negative impact:

Table 4.10 shows that Rsat and Csat are the attributes with the highest negative impact (NI) on MIA_1. These impacts are much higher than the others. Next come the NIs of Psat and Usat. They are also significantly stronger than LS-Tech and D-sat, which in turn are significantly stronger than LS-Sales and LS-Edu.

Insights:

[**Business Insight 19**] The positive impact (PI) ordering of Csat, Psat, Usat, and Dsat was expected. However, the PI of Rsat was surprising because Rsat is frequently considered the most important attribute overall.

[**Known Fact**] The negative impact (NI) ordering was less surprising. Note that Rsat comes in first place. Psat and Usat NI are very similar, and so are LS-Tech and Dsat.

[**Business Insight 20; MF Insight 4**] The most interesting fact in Table 4.10 is that the Fsat attributes have different positive and negative impacts in MIA_1. This information was a novelty for the data manager.

[**Business Insight 21**] The case of Rsat is of special interest. Rsat is constantly rated as the most important attribute. The facts show that its impact is basically a negative one. In other words, reliability is something that is taken as a given by the customers. It has a huge negative impact if it is not there. However, Csat, Psat, Usat, and Dsat all have higher positive impact once high reliability is achieved.

[**Business Insight 22**] The same is true, although at less extent, for LS-Tech (and, we believe for technical support in general).

[**Business Insight 23**] Last but not least, Csat is very important across the board. Capability is a key factor in any scenario.

Diagram group 2: MIA_2

Positive impact:

Table 4.10 shows that Dsat has the highest positive impact (PI) on MIA_2. Next, it comes Csat, Usat, and Psat, all very similar. And last, Rsat with the lowest impact. The local support attributes did not appear in the analysis results.

Negative impact:

Table 4.10 shows that Csat clearly has the strongest negative impact on MIA_2. Next, it comes Psat and Rsat – also with a high negative impact. Usat comes next, and Dsat has the lowest negative impact.

Insights:

[Inconclusive Fact] The results for MIA_2 should be taken very carefully. The distribution of values for the MIA_2 attribute is very different than the distributions for others MIAs. This attribute has a different scale and "not satisfied" (NS) corresponds to [IBM proprietary information removed]. This makes it the only attributes for which $VS\% < NS\%$.

[Inconclusive Fact] By looking at diagram 22, we concluded that Dsat high PI was a "fluke". Although 33% of the customers that were very satisfied with documentation said they were very satisfied with MIA_2, 14.6% of them said they were not satisfied with MIA_2. This number is higher than Csat's 9.6% or Psat's 10.2%. The truth is that all Fsat's positive impacts are similar. Csat would come in first place if we consider together the "very satisfied" and "satisfied" scores.

[Business Insight 24] The NI results are more clear. Csat has the strongest impact and Dsat has the lowest.

[Business Insight 25] MIA_2 is also the only MIA for which Psat's NI is higher than Rsat's NI.

Diagram group 3: MIA_3

Positive impact:

Table 4.10 shows that LS-Sales clearly has the strongest positive impact on MIA_3. Psat, Csat, and LS-Tech come next. And, Rsat and Usat come afterwards. Dsat and LS-Edu diagrams were not generated.

Negative impact:

Table 4.10 shows that Csat, Rsat, and Psat are the attributes with the highest negative impact (NI) on MIA_3. Next, it comes Usat, LS-Sales and LS-Tech.

Insights:

[Business Insight 26] This diagram group has two important pieces of information. The first is the LS-Sales' PI. It is even stronger than the PI on MIA_6. This is probably because sales have a strong influence in [IBM proprietary information removed].

[Business Insight 27] Also note that good local technical support has a sizable positive impact on MIA_6.

[Business Insight 28] The second piece of interesting information is Psat's PI. For the first time, Psat's PI is higher than Csat's PI. This indicates that for some products good Psat is key to MIA_6.

[Known Fact] The negative impacts are not surprising, Csat and Rsat have the strongest NIs on MIA_6. Psat also has a strong NI on MIA_6.

Diagram group 4: MIA_4

Positive impact:

Table 4.10 shows that Csat, Psat, Dsat, Usat are the attributes with the highest positive impact (PI) on MIA_4 (all very similar). Next, it comes LS-Tech and LS-Sales, and last, Rsat and LS-Edu. Also, the PI value distribution for MIA_4 are not as spread as for MIA_1.

Negative impact:

Table 4.10 shows that Rsat clearly has the strongest negative impact on MIA_4. Next, it comes Csat – also with a high negative impact. Usat, Psat, and LS-Tech come in the middle, and Dsat and LS-Sales follows. LS-Edu has the lowest negative impact.

Insights:

[**Business Insight 29**] The impacts of the Fsats on MIA_4 are similar to those on MIA_1. But, there is a very strong difference between the Fsat negative and positive impacts.

[**Business Insight 30**] The difference between Rsat PI and NI in MIA_4 is even stronger than in MIA_1. A high Rsat has one of the lowest positive impacts. A low Rsat has an enormous negative impact in MIA_4. This NI impact seems to imply that low reliability can be very harmful to [IBM proprietary information removed].

[**Business Insight 31**] Csat is again very important across the border.

[**Business Insight 32**] Usat has a strong NI.

Diagram group 5: MIA_5

Positive impact:

Table 4.10 shows that LS-Sales also has the strongest impact on MIA_5. Psat, Usat, and Csat also have strong PIs. And, Rsat come next. Dsat, LS-Edu, and LS-Tech diagrams were not generated by the AF Tool.

Negative impact:

Table 4.10 shows that Rsat clearly has the strongest negative impact on MIA_5. Csat also has a strong impact. Next, it comes Psat and Usat. LS-Sales has a weak negative impact.

Insights:

[**Business Insight 33**] Rsat is again a key factor from the negative impact point of view (NS jumps from 4.7 to 20.7%).

[**Business Insight 34**] Note that Psat and – this time around – Usat also have strong positive impact on MIA_5.

[Business Insight 35] LS-Sales – just like on MIA_3 – has the strongest PI on MIA_5.

Diagram group 6: MIA_6

Positive impact:

Table 4.10 shows that Csat, LS-Sales, and Psat are the attributes with the highest positive impact (PI) on MIA_6. Rsat, Usat, LS-Tech, and LS-Edu come next. And, Dsat has the weakest positive impact.

Negative impact:

Table 4.10 shows that Csat clearly has the strongest negative impact on MIA_6. Next, it comes Rsat, Usat, and Psat – also with a high negative impact. LS-Tech and Dsat come in the middle, and LS-Sales follows. LS-Edu has the lowest negative impact.

Insights:

[Business Insight 36] Csat has the highest PI and NI. This imply that capability is probably the factor that most influences the customer's MIA_6.

[Business Insight 37] Rsat again has a high negative impact. But, this time around, it also has a sizeable positive impact. Rsat, Usat, Psat all seem to be very important to MIA_6.

[Business Insight 38] Dsat on the other hand has the lowest PI. Good Dsat does not seem to have strong influence in MIA_6.

[Business Insight 39] The most interesting fact in this diagram group is the LS-Sales positive impact on MIA_6. Local sales, one of the factors with the lowest NI and PI for MIA_1 and MIA_4, has one of the highest PIs on MIA_6. A hint of what is happening is given by the fact that this impact also happens on MIA_3 and MIA_5. It seems that good sales has a lot of influence on [IBM proprietary information removed]. This shows that, for some MIAs, the local support may be as important as the product factors.

[MF Insight 5] It also shows that Fsats have very different impacts on the MIAs.

[MF Insight 6] This last fact led the data manager to consider that it was important to identify which of the MIAs was the most important for the organization.

[Analyze Further 4] He also said it might be a good idea to run regressions of the MIAs$\times$[CUPRIMDS + local support].

C.5 Results of the the AF Analysis 8

This analysis is described in Section 4.3.5. It involved two attribute classes: "CUPRIMDS × MIAs". These two classes are described in Table 4.11.

This analysis involved only database products. It produced 48 diagrams that were organized in two sets of six diagram groups based on the diagram's explained attribute (MIA_N). The first set of diagrams was ordered by the CUPRIMDS positive impacts on the MIAs. The second set was ordered by the CUPRIMDS negative impacts on the MIAs.

Diagram group 1: MIA_1

Positive impact:

Table 4.12 shows that Psat is the attribute with the highest positive impact (PI) on MIA_1. Usat, Msat, and Csat (in this order) also have a strong PI. Dsat comes next. Isat and Rsat followed by Ssat have the lowest PIs.

Negative impact:

Table 4.12 shows that Csat and Rsat are the attributes with the highest negative impact (NI) on MIA_1. These impacts are much higher than the others. Next comes the NIs of Ssat, Usat, Msat, and Psat. And last comes Isat and Dsat's NI.

Insights:

[**Business Insight 40**] Just like the previous analysis, the positive impact of Rsat is surprisingly low.

[**Business Insight 41**] Ssat PI is also lower than expected. In both cases, their negative impact is very strong.

[**Business Insight 42**] This once again translate to the hypothesis that reliability and technical support are expected as a given by the customers. They are missed dearly when they are absent but they don't have much positive impact when they are present.

[**Business Insight 43**] Also to be noted is the fact that Csat and Rsat's NIs are much stronger than the others. This indicates that if a database is going to miss something it better not be capability or reliability.

[**Known Fact**] Psat had the strongest positive impact on MIA_1. This is not so surprising if you consider we are talking about databases.

[**Business Insight 44**] However, Psat's low NI is somewhat surprising. It probably means that in databases missing performance is less harmful than missing capability, reliability, usability, or maintainability.

[**Business Insight 45**] Usat and Msat's PIs are higher than Csat's. This is somewhat surprising if you compare this analysis with the previous one.

[MF Insight 7; Analyze Further 5] Due to Msat high PIs, the CUST-SAT data manager from the Santa Teresa Lab has raised the hypothesis that the customers are misinterpreting maintainability as the ability to maintain the database instead of the ability to maintain the product. If this is the case, the maintainability question is measuring something different than Msat for DB products. This issue should be further investigated.

Diagram group 2: MIA_2

Positive impact:

Table 4.12 shows that Msat and Psat have the highest positive impacts (PI) on MIA_2. Next, it comes Isat, Usat, and Csat all very similar. Dsat and Rsat come next and Ssat comes last with the lowest impact.

Negative impact:

Table 4.12 shows that Rsat and Csat clearly has the strongest negative impact on MIA_2. Next, it comes Psat followed by Msat and Usat. Isat, Ssat, and Dsat have the lowest negative impact.

Insights:

[Inconclusive Fact] The results for MIA_2 should be taken very carefully. The distribution of values for MIA_2 is very different than the distributions for others MIAs. This attribute has a different scale and "not satisfied" (NS) corresponds to [IBM proprietary information removed]. This makes it the only attribute for which $VS\% < NS\%$.

[Inconclusive Fact] By looking at diagram 22, we concluded that Isat's high PI was a "fluke". Although 29.9% of the customers that were very satisfied with installation said the said they were very satisfied with MIA_2, 16.7% of them said they were not satisfied with MIA_2. This number is higher than Csat's 10.3% or Usat's 12.6%.

[Business Insight 46] Msat and Psat clearly have the highest positive impact, however Msat's PI is a bit inflated by the fact that maintainers give better MIA_2 scores. Csat and Usat would come next (followed by Isat, Rsat, and Ssat), if consider the "very satisfied" and "satisfied answers".

[Business Insight 47] Ssat has a very low PI any way you look at it.

[Business Insight 48] The NI results are more clear. Rsat has the strongest negative impact. Csat also has a strong impact.

[Business Insight 49] Ssat's NI is surprisingly low for MIA_2. May be the service support is not considered by customers that [IBM proprietary information removed] (this might also explain the low PI).

Diagram group 3: MIA_3

Positive impact:

Table 4.12 shows that Ssat and Msat clearly have the strongest positive impact on MIA_3. Usat also has a strong impact. Dsat, Csat, Isat, Psat, and Rsat come next with very similar values.

Negative impact:

Table 4.12 shows that Rsat clearly has the strongest negative impact on MIA_3. Csat comes next. Psat, Dsat, Usat, and Isat follows. Msat and Isat have the weakest NIs.

Insights:

[**Inconclusive Fact**] The biggest surprise is Ssat's PI and NI. This result is a "fluke". It was caused by the fact that people that use the service support have higher MIA_3 scores. Thus, the MIFs distribution changes for the data points used to calculate Ssat PI and NI. PI is (very) inflated and NI is deflated for Ssat.

[**MF Insight 8**] The service support manager told me – during the service support interview – that this was caused by the fact that a person that is [IBM proprietary information removed] tends to cancel or not use the service support. This makes MIA_3 scores higher than average for all Ssat values. This explains the artificially high PI and artificially low NI of Ssat.

[**Inconclusive Fact**] Msat's NI is also artificially inflated (and PI artificially deflated). The same thing that happened with Ssat happens with Msat. The MIFs distribution changes for the data points used to calculate Msat's PI and NI. However, this effect is much smaller for Msat because, while only 30% of the data points have used the service support, 63% of the data points are maintainers.

[**Business Insight 50**] Discounted the effects of Ssat and Msat distorted calculations, probably Usat has the largest PI on MIA_3. It is very interesting that it is higher than Csat and Psat's PI.

[**Business Insight 51**] Also surprising is Psat's low PI and high NI. Psat seems to be a condition necessary but not sufficient for a customer to [IBM proprietary information removed].

Diagram group 4: MIA_4

Positive impact:

Table 4.12 shows that Msat, Usat, Psat are the attributes with the highest positive impact (PI) on MIA_4. Next, it comes Csat, Dsat, and Rsat. Last, it comes Isat and Ssat. Also, the PI value distribution for MIA_4 are not as spread as for MIA_1.

Negative impact:

Table 4.12 shows that – just like for the previous analysis – Rsat clearly has the strongest negative impact on MIA_4. Next, it comes Msat and Csat – also with a high negative impact. Ssat, Usat, and Isat come in the middle. Psat and Dsat have the lowest negative impacts.

Insights:

[**Business Insight 52**] A low Rsat has an enormous negative impact in MIA_4. This NI impact seems to imply that low reliability can be very harmful to [IBM proprietary information removed].

[**Business Insight 53**] Msat seems to be very important across the border with the highest positive impact and the second highest negative impact.

[**Business Insight 54**] Psat case is very interesting. It has a high PI and a very low NI. This may indicate that low performance is not very harmful to [IBM proprietary information removed], but good performance definitely helps it.

[**Business Insight 55**] The same is valid (at less extent) for Usat.

Diagram group 5: MIA_5

Positive impact:

Table 4.12 shows that Msat has the highest PI on MIA_5. Next, it comes Isat, Usat, and Psat. Ssat, Rsat, Csat come in the bottom half. And, Dsat has the weakest positive impact.

Negative impact:

Table 4.12 shows that Csat and Rsat clearly have the strongest negative impact on MIA_5. Psat also has a strong NI. Next, it comes Msat, Usat, and Ssat. Dsat and Isat have the lowest negative impact.

Insights:

[**Business Insight 56**] Csat, Rsat, and Psat are the key factors from the negative impact point of view.

[**Business Insight 57**] Csat has a surprising low PI. We do not have a good explanation for this one.

[**Business Insight 58**] Usat and Psat have very strong PIs.

[**Known Fact**] Msat and Isat's PI are both inflated by the MIA_5 distribution for maintainers and installers. Isat's NI was also deflated a little. Those scores should be looked at with some suspicion.

Diagram group 6: MIA_6

Positive impact:

Table 4.12 shows that Msat, Usat, Psat, and Csat are the attributes with the highest positive impact (PI) on MIA_6. Dsat, Ssat, and Rsat come next. And, Isat has the weakest positive impact.

Negative impact:

Table 4.12 shows that Csat and Rsat clearly have the strongest negative impact on MIA_6. Next, it comes Msat. Usat, Psat and Ssat come in the middle. Isat and Dsat come last with significantly lower negative impacts.

Insights:

[**Business Insight 59**] A low Csat and Rsat have strong negative impacts in MIA_6. This seems to imply that low capability or reliability can be very harmful to [IBM proprietary information removed].

[**Business Insight 60**] Msat seems to be important across the border with the highest positive impact and the third highest negative impact. However, Msat's PI is a bit inflated by the fact that maintainers in general have higher MIA_6.

[**Business Insight 61**] Like before, Psat and Usat have a high PI and medium NI.

Appendix D

Subjective Validation Questionnaire

This appendix presents the questionnaire used to interview the Toronto Lab data manager. The objective of this questionnaire was to subjectively validate the improvement approach presented in this dissertation.

D.1 Questionnaire Introduction

This questionnaire will be used to evaluate the approach we have applied to analyze the CUSTSAT data and its measurement process. Our approach was composed of three steps: measurement characterization, top-down analysis, and bottom-up analysis. The first step – characterization – was executed to identify the data user groups and how they were using the data. The second step – top-down analysis – used GQM to capture the goals of the data users and to map these goals to the metrics and data in the CUSTSAT measurement framework. The third step – bottom-up analysis – used AF to extract knowledge from the CUSTSAT data.

Our evaluation questionnaire contains several multiple choice questions about the three steps of our approach. These questions are designed to characterize the steps that were useful AS WELL AS those that were not with respect to different parts of the measurement process. Please, be as accurate as possible in your answers.

Our evaluation questionnaire also has some multiple choice questions referring to the existing CUSTSAT measurement framework. Those questions refer to the current processes, methods, and mechanisms utilized to measure, maintain, and analyze the CUSTSAT data, as well as to improve the CUSTSAT survey questionnaire. These questions are necessary to characterize what type of new capabilities our approach has added to the existing CUSTSAT measurement framework.

Some of the multiple choice questions are followed by open ended questions of the type "please, justify your answer". These questions will be used to expand on the issues addressed by the multiple choice questions. They should be answered verbally. We will use a tape recorder to record the answers. They are very important to us.

We use the following acronyms throughout the evaluation questionnaire:

- GQM: goal-question-metric.
- AF: attribute focusing.
- MC: measurement characterization.
- CUSTSAT: customer satisfaction.
- CUSTSAT MF: the existing CUSTSAT measurement framework.
- Toronto Lab: IBM Toronto Laboratory.
- DA/P: data analysis or data presentation.
- QQ: CUSTSAT questionnaire question.

Please, do not hesitate to ask for clarification or explanations during this interview.

D.2 Questionnaire

Part 1 - Knowledge discovery and data visualization.

Let us start by talking about knowledge discovery:

Q1.1.1 *[O1,G1]* - You have used the bottom-up method (AF Analyses) as a knowledge discovery mechanism in your database. After this experience, tell us how important you think intelligent data exploration and knowledge discovery mechanisms are to your business ?

(A) No importance at all.

(B) Little importance - the traditional statistical methods already produce most of the information that I need. These methods can only help us to extract a little extra information from the data we have.

(C) Some importance - these methods can complement the traditional statistical methods by providing some new and interesting information for me.

(D) Great importance - these methods will have a sizable impact in our business; only with them we can take full advantage of the data we have.

(E) Absolute importance - these methods must be applied in our line of business.

Q1.1.2 *[O1,G2.1]* - Prior to the use of AF, how good was the existing CUSTSAT MF in providing tools and methods for intelligent data exploration and knowledge discovery in your database ?

(A) Very Poor - no support was provided

(B) Poor - the framework did not have adequate mechanisms for this type of data exploration, however we could use our tools to manually do some intelligent data exploration.

(C) Fair - there were some tools and methods for knowledge discovery but much improvement was still needed.

(D) Good - there existed adequate tools and methods for knowledge discovery, but there still was room for improvement.

(E) Very good - our knowledge discovery support was state-of-the-art, very little improvement was possible.

Q1.1.2.1 *[O1,G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q1.1.3 *[O1,G2.2]*- How good was the bottom-up method (AF Analysis) in discovering knowledge (interesting and unknown information) in your database ?

(A) Very Poor - no useful information was discovered

(B) Poor - the discovered information was of little use and interest to our business

(C) Fair - there were some interesting facts discovered but they had a limited impact on our business.

(D) Good - there were interesting facts discovered that had a significant impact on our business.

(E) Very good - there were interesting facts discovered that had a major impact on our business.

Q1.1.3.1 *[O1,G2.2]* - Please, justify your answer to the previous question.

Let us now talk about data visualization:

Q1.2.1 *[O2,G1]* - How important do you think visualization of data distributions and associations is to your business ?

(A) No importance at all.

(B) Little importance - simple tables and percentages are sufficient in most cases.

(C) Some importance - visualization can complement the percentages and tables by making the data easier to interpret.

(D) Great importance - we need visualization to interpret some important data distributions and associations.

(E) Absolute importance - we must use visualization, without it we cannot interpret the data we have.

Q1.2.2 *[O2,G2.1]* - How good was the existing CUSTSAT MF in providing tools and methods for visualizing the data distributions and associations you have ?

(A) Very Poor - no support was provided

(B) Poor - the framework did not have adequate mechanisms for visualization, however we could use our numerical data to manually build diagrams and charts.

(C) Regular - there was some tools and methods for data visualization, but much improvement was still needed.

(D) Good - there existed adequate tools and methods for data visualization, but there still was room for improvement.

(E) Very good - our data visualization support was state-of-the-art, very little improvement was possible.

Q1.2.2.1 *[O2,G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q1.2.3 *[O2,G2.2]* - Considering only those AF diagrams with potentially interesting information, how good was the bottom-up method (AF Analysis) in helping you to visualize data distributions and associations ?

(A) Very Poor - no useful diagrams was produced.

(B) Poor - the produced diagrams were of little use to me, they were very difficult to interpret.

(C) Fair - there were useful diagrams produced, but they were not easy to interpret.

(D) Good - the tool produced useful and easy to interpret diagrams, but there are still needed improvements.

(E) Very good - the tool produced useful, complete, and easy to interpret diagrams.

Q1.2.3.1 *[O2,G2.1]* - Please, justify your answer to the previous question.

Part 2 - Evaluating questions (QQs) and questionnaire format

Let us talk about the format and wording of QQs.

Q2.1.1 *[O4,G1]* - How important is it for you to have methods/processes to assess the adequacy of the wording and structure (scale and possible values) of the QQs ?

(A) No importance at all.

(B) Little importance - the QQs are stable and the QQ wording leaves little room for misunderstanding by the interviewed subjects.

(C) Some importance - QQs are subject to small reviews and their format has some impact on the answers we obtain.

(D) Great importance - QQs are subject to periodic reviews and their format has significant impact on the answers that we obtain.

(E) Absolute importance - QQs are frequently modified and their format has strong impact on the answers that we obtain.

Q2.1.2 *[O4,G2.1]* - How good is the existing CUSTSAT MF process in assessing and reviewing the wording and structure (scale and possible values) of the QQs?

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for assessing and reviewing QQs.

(C) Fair - there is a process for assessing and reviewing QQs, but much improvement is still needed.

(D) Good - there is an adequate process for assessing and reviewing QQs, but there still is room for improvement.

(E) Very good - there is a successful process for assessing and reviewing QQs, very little improvement is possible.

Q2.1.2.1 *[O4,G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q2.1.3 thru 6 - How helpful were each of the steps of the method in assessing the wording and structure (scale and possible choices/values) of the QQs ?

(A) Useless - the step was not helpful at all for this task.

(B) Helped a little - it helped very little our assessment of the wording and structure of the QQs.

(C) Helped somewhat - in some cases, it helped our assessment of the wording and structure of the QQs.

(D) Helped significantly - it had a significant impact on our assessment of the wording and structure of the QQs.

(E) Helped a lot - it had a major impact on our assessment of the wording and structure of the QQs.

Q2.1.3 *[O4;G2.2]* - The MC step: A B C D E

Q2.1.4 *[O4;G2.2]* - The top-down (GQM) analysis: A B C D E

Q2.1.5 *[O4,G2.2]* - The bottom-up (AF) analysis: A B C D E

Q2.1.6 *[O4,G2.2]* - Please, justify your answer to the previous questions.

Now let us talk about the questionnaire organization.

Q2.2.1 *[O5,G1]* - How important is it for you to have methods/processes for assessing and reviewing the questionnaire format (i.e. QQs dependency, ordering, and grouping inside the questionnaire) ?

(A) No importance at all.

(B) Little importance - the questionnaire format is stable and very few modifications are made to it.

(C) Some importance - the questionnaire format is stable but some periodic assessment is needed to include/exclude a few QQs.

(D) Great importance - the questionnaire format is subject to some modifications; we periodically have to review the organization of the QQs, and include/exclude some QQs.

(E) Absolute importance - the questionnaire format is subject to major modifications, we periodically have to review the organization of the QQs, and include/exclude many QQs.

Q2.2.2 *[O5;G2.1]* - How good is the existing CUSTSAT MF process in assessing and reviewing the questionnaire format (i.e. QQs dependency, ordering, and grouping inside the questionnaire) ?

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for assessing and reviewing the questionnaire format.

(C) Fair - there is a process for assessing and reviewing the questionnaire format, but much improvement is still needed.

(D) Good - there is an adequate process for assessing and reviewing the questionnaire format, but there still is room for improvement.

(E) Very good - there is a successful process for assessing and reviewing the questionnaire format, very little improvement is needed.

Q2.2.2.1 *[O5;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q2.2.3 thru 6 - How helpful were each of the steps in assessing the questionnaire format (i.e. QQs dependency, ordering, and grouping inside the questionnaire) ?

(A) Useless - the step was not helpful at all for this task.

(B) Helped a little - it helped very little our assessment of the questionnaire format.

(C) Helped somewhat - in some cases, it helped our assessment of the questionnaire format.

(D) Helped significantly - it had a significant impact on our assessment of the questionnaire format.

(E) Helped a lot - it had a major impact on our assessment of the questionnaire format.

Q2.1.3 *[O5;G2.2]* - The MC step: A B C D E

Q2.1.4 *[O5;G2.2]* - The top-down (GQM) analysis: A B C D E

Q2.1.5 *[O5;G2.2]* - The bottom-up (AF) analysis: A B C D E

Q2.1.6 *[O5;G2.2]* - Please, justify your answer to the previous questions.

Part 3 - Assessing the importance of questions (QQs), data analyses and data presentations (DA/Ps)

Let us now talk about the usefulness of QQs:

Q3.1.1 thru 3 - How important is it to have methods/processes for assessing the usefulness of the QQs to:

Q3.1.1 *[O3,G1]* - specific user groups ? A B C D E

Q3.1.2 *[O3,G1]* - the organizations as a whole ? A B C D E

(A) No importance at all.

(B) Little importance - we usually know how important each QQ is.

(C) Some importance - sometimes, we need to re-assess the usefulness of QQs (e.g., due to questionnaire modifications, user group re-organizations, or changes in the needs of user groups).

(D) Great importance - we need to constantly re-assess the usefulness of QQs (e.g., due to periodic questionnaire modifications, user group re-organizations, or changes in the needs of user groups).

(E) Absolute importance - we must continuously re-assess the usefulness of QQs (e.g., due to frequent questionnaire modifications, user group re-organizations, or changes in the needs of user groups).

Q3.1.3 *[O3;G1]* - If you chose C, D, E to the previous questions, explain why you need to (re-)assess the importance of QQs.

Q3.1.4 thru 6 - How good is the CUSTSAT MF process for assessing the usefulness of the QQs to:

Q3.1.4 *[O3;G2.1]* - specific user groups ? A B C D E

Q3.1.5 *[O3;G2.1]* - the organizations as a whole ? A B C D E

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for assessing the importance of QQs.

(C) Fair - there is a process for assessing the importance of QQs, but much improvement is still needed.

(D) Good - there is an adequate process for assessing the importance of QQs, but there still is room for improvement.

(E) Very good - there is a successful process for assessing the importance of QQs, very little improvement is needed.

Q3.1.6 *[O3;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous questions.

Q3.1.7 thru 10 - How helpful was each step of the new approach in assessing the importance of the QQs to specific user groups ?

(A) Useless - the step was not helpful at all for this task.

(B) Helped a little - it was of limited help in assessing the importance of a few QQs to some user groups.

(C) Helped somewhat - in some cases, it helped our assessment of the importance of some QQs to some user groups.

(D) Helped significantly - it helped our assessment of the importance of several QQs to several user groups.

(E) Helped a lot - it had a major impact on our assessment of the importance of the QQs to most of the user groups.

Q3.1.7 *[O3;G2.2]* - The MC step: A B C D E

Q3.1.8 *[O3;G2.2]* - The top-down (GQM) analysis: A B C D E

Q3.1.9 *[O3;G2.2]* - The bottom-up (AF) analysis: A B C D E

Q3.1.10 *[O3;G2.2]* - Please, justify your answers to the previous questions.

Q3.1.11 thru 14 - How helpful was each step of the new approach in assessing the importance of the QQs to the organization as a whole ?

(A) Useless - the MC step was not helpful at all for this task.

(B) Helped a little - it was of limited help in assessing the overall importance of a few QQs.

(C) Helped somewhat - in some cases, it helped our assessment of the overall importance of some QQs.

(D) Helped significantly - it helped our assessment of the overall importance of several QQs.

(E) Helped a lot - it had a major impact on our assessment of the QQs importance overall.

Q3.1.11 *[O3;G2.2]* - The MC step: A B C D E

Q3.1.12 *[O3;G2.2]* - The top-down (GQM) analysis: A B C D E

Q3.1.13 *[O3;G2.2]* - The bottom-up (AF) analysis: A B C D E

Q3.1.14 *[O3;G2.2]* - Please, justify your answers to the previous questions.

Let us now talk about assessing the usefulness of data analyses and presentations (DA/Ps):

Q3.2.1 thru 3 - How important is it to have methods/processes for assessing the usefulness of the DA/Ps to:

Q3.2.1 *[O6;G1]* - specific user groups ? A B C D E

Q3.2.2 *[O6;G1]* - the organizations as a whole ? A B C D E

(A) No importance at all.

(B) Little importance - we usually know how important each DA/P is.

(C) Some importance - sometimes we need to re-assess the usefulness of DA/Ps (e.g., due to questionnaire modifications, user group re-organizations, or changes in the needs of the user groups or the organization).

(D) Great importance - we need to constantly re-assess the usefulness of DA/Ps (e.g., due to periodic questionnaire modifications, user group re-organizations, or changes in the needs of the user groups or the organization).

(E) Absolute importance - we must continuously re-assess the usefulness of DA/Ps (e.g., due to frequent questionnaire modifications, user group re-organizations, or changes in the needs of the user groups or the organization).

Q3.2.3 *[O6;G1]* - If you chose C, D, or E to the previous questions, explain why you need to (re-)assess the importance of DA/Ps.

Q3.2.4 thru 6 - How good is the CUSTSAT MF process for assessing the usefulness of the DA/Ps to:

Q3.2.4 *[O6;G2.1]* - specific user groups ? A B C D E

Q3.2.5 *[O6;G2.1]* - the organizations as a whole ? A B C D E

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for assessing the importance of DA/Ps.

(C) Fair - there is a process for assessing the importance of DA/Ps, but much improvement is still needed.

(D) Good - there is an adequate process for assessing the importance of DA/Ps, but there still is room for improvement.

(E) Very good - there is a successful process for assessing the importance of DA/Ps, very little improvement is needed.

Q3.2.6 *[O6;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous questions.

Q3.2.7 thru 10 - How helpful was each step of the new approach in assessing the importance of the DA/Ps for specific user groups ?

(A) Useless - the step was not helpful at all for this task.

(B) Helped a little - it was of limited help in assessing the importance of a few DA/Ps to some user groups.

(C) Helped somewhat - in some cases, it helped our assessment of the importance of some DA/Ps to some user groups.

(D) Helped significantly - it helped our assessment of the importance of several DA/Ps to several user groups.

(E) Helped a lot - it had a major impact on our assessment of the DA/Ps importance to most of the user groups.

Q3.2.7 *[O6;G2.2]* - The MC step: A B C D E

Q3.2.8 *[O6;G2.2]* - The top-down (GQM) analysis: A B C D E

Q3.2.9 *[O6;G2.2]* - The bottom-up (AF) analysis: A B C D E

Q3.2.10 *[O6;G2.2]* - Please, justify your answers to the previous questions.

Q3.2.11 thru 14 - How helpful was each step of the new approach in assessing the overall importance of the DA/Ps for the organization ?

(A) Useless - the step was not helpful at all for this task.

(B) Helped a little - it was of limited help in assessing the overall importance of a few DA/Ps.

(C) Helped somewhat - in some cases, it helped our assessment of the overall importance of some DA/Ps.

(D) Helped significantly - it helped our assessment of the overall importance of several DA/Ps.

(E) Helped a lot - it had a major impact on our assessment of the overall importance of the DA/Ps.

Q3.2.11 *[O6;G2.2]* - The MC step: A B C D E

Q3.2.12 *[O6;G2.2]* - The top-down (GQM) analysis: A B C D E

Q3.2.13 *[O6;G2.2]* - The bottom-up (AF) analysis: A B C D E

Q3.2.14 *[O6;G2.2]* - Please, justify your answers to the previous questions.

Part 4 - Understanding user needs and goals

Let us now talk about the goals of data users (i.e. the objectives that Toronto Lab groups want to achieve by using the CUSTSAT data):

Q4.1.1 *[O8;G1]* - How important is it to have methods/processes to understand the goals of CUSTSAT data users ?

(A) No importance at all.

(B) Little importance - we know the needs of (current and prospective) CUSTSAT data users well and they do not change much over time.

(C) Some importance - sometimes we need to re-assess the needs of data users, due to occasional changes in the organization, consumer market, products, and other factors affecting them.

(D) Great importance - we need to periodically re-assess the needs of data users, due to periodic changes in the organization, consumer market, products, and other factors affecting them.

(E) Absolute importance - we must continuously re-assess the needs of data users, due to frequent changes in the organization, consumer market, products, and other important factors affecting them.

Q4.1.1.1 *[O8;G1]* - If you chose C, D, or E, explain what factors affects the needs of users with respect to the CUSTSAT data. Otherwise skip to question 4.1.4.

Q4.1.2 *[O8;G2.1]* - How good is the CUSTSAT MF process in understanding the goals of data users ?

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for assessing the the goals of data users.

(C) Fair - there is a process for assessing the goals of users, but much improvement is still needed.

(D) Good - there is an adequate process for assessing the goals of users, but there still is room for improvement.

(E) Very good - there is a successful process for assessing the goals of users, very little improvement is needed.

Q4.1.2.1 *[O8;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q4.1.3 *[O7;G2.1]* - How good is the CUSTSAT MF process in mapping the goals of data users to their needs with respect to data collection (QQ), analysis and presentation (DA/P) ?

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for mapping the goals of data users to the QQs and DA/Ps.

(C) Fair - there is a process for mapping the data user goals to the QQs and DA/Ps, but much improvement is still needed.

(D) Good - there is an adequate process for mapping the data user goals to the QQ and DA/Ps, but there still is room for improvement.

(E) Very good - there is a successful process for mapping the data user goals to the QQs and DA/Ps, very little improvement is needed.

Q4.1.3.1 *[O7;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

Q4.1.4 *[O8;G1]* - How important is it to have a mechanism to document the goals and needs of data users ?

(A) No importance at all.

(B) Little importance - we know the needs of the data users well; formal documentation would be of marginal use.

(C) Some importance - formal documentation would be of some use as a channel of communication with the user groups or as an input to the questionnaire improvement process.

(D) Great importance - formal documentation can work as an important communication channel with the user groups or as an important input to the questionnaire improvement process.

(E) Absolute importance - formal documentation is a fundamental communication channel with the user groups or a fundamental input to the questionnaire improvement process.

Q4.1.4.1 *[O8;G1]* - If you chose C, D, or E, explain how the documentation of the needs of data users would help you. Otherwise skip to question 4.1.6.

Q4.1.5 *[O8;G2.1]* - How good is the CUSTSAT MF process in documenting the needs of the CUSTSAT data users ?

(A) Very Poor - there is no such process.

(B) Poor - the framework has an inadequate process for documenting the needs of the data users.

(C) Fair - there is a process for documenting the needs of users, but much improvement is still needed.

(D) Good - there is an adequate process for documenting the needs of users, but there still is room for improvement.

(E) Very good - there is a successful process for documenting the user needs, very little improvement is needed.

Q4.1.5.1 *[O8;G2.1]* - Please explain what kinds of improvements are needed, if you have answered B, C, or D to the previous question.

The measurement characterization (MC) step is used to document the metrics (QQ), data users, and how they are using the data.

Q4.1.6 *[O7;G2.2]* - How helpful was the MC step in better understanding the needs of data users with respect to QQs and DA/Ps ?

(A) Useless - the MC step did not help us at all to better understand the needs of data users.

(B) Helped a little - it was of limited help in better understanding needs of data users.

(C) Helped somewhat - in some cases, it helped us to better understand data user needs.

(D) Helped significantly - it helped us to better understand the needs of several user groups.

(E) Helped a lot - it had a major impact on our understanding of the needs of data user groups.

Q4.1.6.1 *[O7;G2.2]* - Please, justify your answer to the previous question.

The GQM-based method tries to understand the needs of data users by capturing their goals, and by documenting and mapping them to the QQs and DA/Ps.

Q4.1.7 *[O7;G2.2]* - How helpful was the top-down (GQM) analysis in improving your understanding of the needs of CUSTSAT data users with respect to QQs and DA/Ps ?

(A) Useless - the top-down (GQM) analysis did not help us at all to better understand the needs of data users.

(B) Helped a little - it was of limited help in better understanding the needs of data users.

(C) Helped somewhat - in some cases, it helped us to better understand the needs of data users.

(D) Helped significantly - it helped us to better understand the needs of several user groups.

(E) Helped a lot - it had a major impact on our understanding of the needs of data user groups.

Q4.1.7.1 *[O7;G2.2]* - Please justify your answer to the previous question. Please skip to Q4.1.11, if you chose A in the previous question.

Q4.1.8 *[O8;G2.2]* - How successful is the top-down (GQM) analysis in capturing the goals of data users ?

(A) Very Poor - the analysis does not capture the data user goals.

(B) Poor - the analysis is inadequate for capturing the goals of data users.

(C) Fair - the analysis is adequate for capturing the goals of data users, but much improvement is still needed.

(D) Good - the analysis is adequate for capturing the goals of data users, but there still is room for improvement.

(E) Very good - the analysis is successful in capturing the goals of data users, very little improvement is needed.

Q4.1.8.1 *[O8;G2.2]* - Please justify your answer to the previous question.

Q4.1.9 *[O8;G2.2]* - How good was the top-down (GQM) analysis in mapping the goals of data users to their needs with respect to QQs and DA/Ps ?

(A) Very Poor - the mapping process does not work at all.

(B) Poor - the process for mapping the goals of data users to the QQs and DA/Ps is deficient in several aspects.

(C) Fair - the process for mapping the goals of data users to the QQs and/or DA/Ps is adequate, but much improvement is still needed.

(D) Good - the process for mapping the goals of data users to the QQ and DA/Ps is adequate, but there still is room for improvement.

(E) Very good - the process for mapping the goals of data users to the QQs and DA/Ps fulfills our requirements, very little improvement is needed.

Q4.1.9.1 *[O8;G2.2]* - Please justify your answer to the previous question.

Q4.1.10 *[O8;G2.2]* - How good is the top-down (GQM) analysis in documenting the needs of data users ?

(A) Very Poor - this type of documentation is not useful at all.

(B) Poor - this type of documentation has little value to us (it is inadequate for our needs).

(C) Fair - this type of documentation has limited value to us (it is adequate for our needs but requires much improvement).

(D) Good - this type of documentation has significant value to us (it is adequate for our needs, but requires some improvement).

(E) Very good - this type of documentation has great value to us (it fits our needs very well).

Q4.1.10.1 *[O8;G2.2]* - Please justify your answer.

Part 5 - Identifying new applications and user groups for the measured data, and defining new questions (QQs), data analyses, and presentations (DA/Ps).

Let us talk about identifying new applications and user groups for the CUSTSAT data.

Q5.1.1 *[O9;G1]* - How important is it to have mechanisms for identifying new applications and user groups for the measured data ?

(A) No importance at all.

(B) Little importance - we know the (current and prospective) CUSTSAT data user groups in our organization well and how they use the data.

(C) Some importance - sometimes we need to re-assess the usefulness of our data for new user groups inside the organization.

(D) Great importance - we need to periodically re-assess the usefulness of our data for new user groups inside the organization.

(E) Absolute importance - we must continuously re-assess the usefulness of our data for new user groups inside the organization.

Q5.1.2 *[O9;G2.1]* - How good are the CUSTSAT MF mechanisms in identifying new applications and user groups for the CUSTSAT data ?

(A) Very Poor - there are no such mechanisms.

(B) Poor - there are a few mechanisms, but they are ineffective.

(C) Fair - there are some mechanisms, but we need much improvement in this aspect.

(D) Good - there are adequate mechanisms, but we still need to improve them.

(E) Very good - there are successful mechanisms for identifying new and prospective data user groups, very little improvement is needed.

Q5.1.2.1 *[O9;G2.1]* - Please justify your answer to the previous question.

Q5.1.3 *[O9;G2.2]* - How helpful was the MC step in identifying new or prospective applications and user groups for the CUSTSAT data ?

(A) Useless - the MC step was not helpful at all for this task.

(B) Helped a little - it was of limited help in identifying new or prospective applications and user groups for the CUSTSAT data.

(C) Helped somewhat - it helped us to identify a few new or prospective applications and user groups.

(D) Helped significantly - it helped us to identify some new or prospective applications and user groups.

(E) Helped a lot - it helped us to identify several new or prospective applications and user groups.

Q5.1.4 *[O9;G2.2]* - How about the bottom-up (AF) analysis ? A B C D E

Q5.1.5 *[O9;G2.2]* - Please, justify your answer to the previous questions.

Let us now talk about defining new questions for the CUSTSAT questionnaire (QQs).

Q5.2.1 *[O10;G1]* - How important is it to have mechanisms for defining new QQs ?

(A) No importance at all.

(B) Little importance - our questionnaire is complete and needs little updating.

(C) Some importance - sometimes we need to update the questionnaire based on occasional changes in the market, our organization, or the needs of data users.

(D) Great importance - we need to periodically update the questionnaire based on frequent changes in the market, our organization, or the needs of data users.

(E) Absolute importance - we must continuously update the questionnaire based on constant changes in the market, our organization, or the needs of data users.

Q5.2.2 *[O10;G2.1]* - How good are the CUSTSAT MF mechanisms for defining new QQs ?

(A) Very Poor - there are no such mechanisms.

(B) Poor - there are such mechanisms, but they are ineffective.

(C) Fair - there are such mechanisms, but we need much improvement.

(D) Good - there are adequate mechanisms, but we still need to improve them.

(E) Very good - there are successful mechanisms for defining new QQs, very little improvement is need.

Q5.2.2.1 *[O10;G2.1]* - Please justify your answer to the previous question.

Q5.2.3 *[O10;G2.2]* - How helpful was the bottom-up (AF) analysis in defining new QQs ?

(A) Useless - the bottom-up (AF) analysis was not helpful at all for this task.

(B) Helped a little - in some very limited cases, it pointed to new types of information we needed to collect.

(C) Helped somewhat - in a few cases, it pointed to new types of information we needed to collect.

(D) Helped significantly - in some cases, it pointed to new types of information we needed to collect.

(E) Helped a lot - in many cases, it pointed to new types of information we needed to collect.

Q5.2.4 *[O10;G2.2]* - How about the top-down (GQM) analysis ? A B C D E

Q5.2.5 *[O10;G2.2]* - Please, justify your answer to the previous questions.

Let us now talk about defining new data presentations and data analyses – DA/Ps.

Q5.3.1 *[O10;G1]* - How important is it to have mechanisms for defining new DA/Ps ?

(A) No importance at all.

(B) Little importance - our DA/Ps are complete and need little updating.

(C) Some importance - sometimes we need to update the DA/Ps based on occasional changes in the market, our organization, or the needs of data users.

(D) Great importance - we need to periodically update the DA/Ps based on constant changes in the market, our organization, or the needs of data users.

(E) Absolute importance - we must continuously update the DA/Ps based on frequent changes in the market, our organization, or the needs of data users.

Q5.3.2 *[O10;G2.1]* - How good is the CUSTSAT MF mechanisms for defining new DA/Ps ?

(A) Very Poor - there are no such mechanisms.

(B) Poor - there are such mechanisms, but they are ineffective.

(C) Fair - there are such mechanisms, but we need much improvement.

(D) Good - there are adequate mechanisms, but we still need to improve them.

(E) Very good - there are successful mechanisms to define new DA/Ps, very little improvement is needed.

Q5.3.2.1 *[O10;G2.1]* - Please justify your answer to the previous question.

Q5.3.3 *[O10;G2.2]* - How helpful was the bottom-up (AF) analysis in defining new DA/Ps ?

(A) Useless - the bottom-up (AF) analysis was not helpful at all for this task.

(B) Helped a little - in some very limited cases, it pointed to new types of information we needed to study more closely.

(C) Helped somewhat - in a few cases, it pointed to new types of information we needed to study more closely.

(D) Helped significantly - in some cases, it pointed to new types of information we needed to study more closely.

(E) Helped a lot - in many cases, it pointed to new types of information we needed to to study more closely.

Q5.3.4 *[O10;G2.2]* - How about the top-down (GQM) analysis ? A B C D E

Q5.3.5 *[O10;G2.2]* - Please, justify your answer to the previous questions.

Part 6 - Evaluating the cost effectiveness of the approach

In order to conclude, let us talk about the cost effectiveness of the steps that composed our approach.

Q6.1.1 *[G3]* - Considering your experiences with our approach and your answers during this interview, would you say that applying the MC step to the CUSTSAT MF was:

(A) Of no value - the benefits of the method definitely were not worth the resources we spent to apply it.

(B) Of meagre value - it produced some benefits but they probably did not outweigh the resources we spent applying it.

(C) Of modest value - it had some benefits but also considerable cost, and I don't know which outweighs the other.

(D) Of some value - its benefits probably outweigh the resources we spent applying it.

(E) Of considerable value - its benefits definitely outweigh the resources spent to apply the method.

Q6.1.2 *[G3]* - In your opinion, what were the main benefits of the MC step ?

Q6.1.3 *[G3]* - In your opinion, what were the main drawbacks of the MC step ?

Q6.2.1 *[G3]* - How about the top-down (GQM) analysis (use same scale as before):
A B C D E

Q6.2.2 *[G3]* - In your opinion, what were the main benefits of the top-down (GQM) analysis ?

Q6.2.3 *[G3]* - In your opinion, what were the main drawbacks of the top-down (GQM) analysis ?

Q6.3.1 *[G3]* - How about the bottom-up (AF) analysis (use same scale as before):
A B C D E

Q6.3.2 *[G3]* - In your opinion, what were the main benefits of the bottom-up (AF) analysis ?

Q6.3.3 *[G3]* - In your opinion, what were the main drawbacks of the bottom-up (AF) analysis ?

Bibliography

[1] Y. Akao, editor. *Quality Function Deployment: Integrating Customer Requirements Into Product Design.* Productivity Press, Cambridge MA, 1987.

[2] A. J. Albrech and J. E. Gaffney. Software function, source lines of code, and development effort prediction: a software science validation. *IEEE Transactions on Software Eng.*, 9(6):639–647, November 1983.

[3] L. J. Arthur. Quantum improvements in software system quality. *Communications of the ACM*, 40(6):47–52, June 1997.

[4] J. Barnard and A. Price. Managing code inspection information. *IEEE Software*, 11(2):59–69, March 1994.

[5] K. M. Bartol and D. C. Martin. *Management*, chapter 7. McGraw Hill Series in Management. McGraw Hill, 1991.

[6] V. R. Basili. The Experience Factory: Can it make a 5 ? In *Proceedings of the Seventeenth Annual Software Engineering Workshop*, number SEL-92-004 in Software Engineering Laboratory Series, Greenbelt MD, December 1992.

[7] V. R. Basili, G. Caldiera, and H. D. Rombach. The Experience Factory. In *Encyclopedia of Software Engineering*, pages 469–476. John Wiley & Sons, 1994.

[8] V. R. Basili, M. K. Daskalantonakis, and R. H. Yacobellis. Technology transfer at Motorola. *IEEE Software*, 11(2):70–76, March 1994.

[9] V. R. Basili and R. W. Reiter Jr. Evaluating automatable measures of software development. In *Workshop on Quantitative Software Models*, Kiamesha NY, October 1979. IEEE.

[10] Victor R. Basili. *Models and Metrics for Software Management and Engineering.* IEEE Tutorial. IEEE Computer Society Press, Los Alamitos CA, 1980.

[11] Victor R. Basili. Can we measure software technology: Lessons learned from 8 years trying. In *Proc. 10th Annual Software Engineering Workshop*, Greenbelt MD, 1985. NASA/GSFC.

[12] Victor R. Basili. Quantitative evaluation of software engineering methodology. In *Proc. of the First Pan Pacific Computer Conference*, Melbourne Australia, September 1985.

[13] Victor R. Basili. Applying the Goal/Question/Metric paradigm in the Experience Factory. In *10th Annual CSR Workshop*, October 1993.

[14] Victor R. Basili. The Experience Factory and its relationship to other quality approaches. *Advances in Computers*, 41(1):65–82, 1995.

[15] Victor R. Basili and S. Green. Software process evolution at the SEL. *IEEE Software*, 11(4):58–66, July 1994.

[16] Victor R. Basili and D. H. Hutchens. An empirical study of a syntactic complexity family. *IEEE Transactions on Software Eng.*, 9(6):664–672, November 1983.

[17] Victor R. Basili and H. D. Rombach. The TAME project: Towards improvement-oriented software environments. *IEEE Transactions on Software Eng.*, 14(6):758–773, June 1988.

[18] Victor R. Basili and R. W. Selby. Paradigms for experimentation and empirical studies in software engineering. *Reliability Engineering and System Safety*, 32:171–191, 1991.

[19] Victor R. Basili, R. W. Selby, and T. Y. Phillips. Metric analysis and validation across FORTRAN projects. *IEEE Transactions on Software Eng.*, 9(6):652–663, November 1983.

[20] Victor R. Basili and D. M. Weiss. A methodology for collecting valid software engineering data. *IEEE Transactions on Software Eng.*, 10(6):728–738, November 1984.

[21] I. S. Bhandari. Attribute focusing: Machine-assisted knowledge discovery applied to software production process control. *Knowledge Acquisition Journal*, 6(3):271–294, September 1994.

[22] I. S. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap, and K. Ramanujam. Advanced scout: Data mining and knowledge discovery in the NBA data. *Data Mining and Knowledge Discovery*, 1(1):121–125, January 1997.

[23] I. S. Bhandari, M. J. Halliday, J. Chaar, R. Chillarege, K. Jones, J. S. Atkinson, C. Lepori-Costello, P. Y. Jasper, E.D. Tarver, C. C. Lewis, and M. Yonezawa. In-process improvement through defect data interpretation. *IBM Systems Journal*, 33(1), January 1994.

[24] I. S. Bhandari, M. J. Halliday, E. Tarver, D. Brown, J. Chaar, and R. Chillarege. A case study of software process improvement during development. *IEEE Transactions on Software Eng.*, 19(12):1157–1170, December 1993.

[25] I. S. Bhandari, M. G. Mendonça, and J. Dawson. On the use of machine-assisted knowledge discovery to analyze and reengineer measurement frameworks. In *Proc. of CASCON'95*, Toronto ON, November 1995.

[26] I. S. Bhandari, B. Ray, M. Y. Wong, D. Choi, A. Watanabe, R. Chillarege, M. Halliday, A. Dooley, and J. Chaar. An inference structure for process feedback: Technique and implementation. *Software Quality Journal*, 3(3):167–189, 1994.

[27] B. W. Boehm. Software engineering economics. *IEEE Transactions on Software Eng.*, 10(1):4–21, January 1984.

[28] B. W. Boehm, J. R. Brown, and M. Lipow. Quantitative evaluation of software quality. In *2nd International Conference on Software Engineering*, pages 592–605, San Francisco CA, October 1976. IEEE & ACM.

[29] Lionel C. Briand, Victor R. Basili, and Christopher Hetmanski. Developing interpretable models with optimized set reduction for identifying high-risk software components. *IEEE Transactions on Software Eng.*, 19(11):1028–1044, November 1993.

[30] Lionel C. Briand, Victor R. Basili, and Sandro Morasca. Goal-driven definition of product metrics based on properties. Technical Report CS-TR 3346 / UMIACS-TR 94-106, University of Maryland, College Park MD, 1994.

[31] Lionel C. Briand, Victor R. Basili, and Sandro Morasca. Property-based software engineering measurement. Technical Report CS-TR 3368 / UMIACS-TR 94-75, University of Maryland, College Park MD, 1994.

[32] Lionel C. Briand, Victor R. Basili, and W. M. Thomas. A pattern recognition approach for software engineering data analysis. *IEEE Transactions on Software Eng.*, 18(11):931–942, November 1992.

[33] F. P. Brooks. *The Mythical Man-Month: Essays on Software Engineering*. Addison-Wesley Publishing Company, Reading MA, 2nd edition, July 1978.

[34] F. P. Brooks. No silver bullet: Essence and accidents of software engineering. *IEEE Computer*, 20(10):19, 1987.

[35] B. G. Buchanan and T. M. Mitchell. Model-directed learning of production rules. In *Pattern Directed Inference Systems*. Academic Press, 1978.

[36] J. G. Carbonell, R. S. Michalski, and T. M Mitchell. An overview of machine learning. In J. G. Carbonell, R. S. Michalski, and T. M Mitchell, editors, *Machine Learning, an Artificial Intelligence Approach*, volume 1, pages 3–24. Morgan Kaufmann, San Mateo CA, 1983.

[37] E. Colet and I. S. Bhandari. Statistical issues in the application of data mining to the NBA using attribute focusing. In *Proceedings of the 1997 Joint Statistical Meetings*, Anahein CA, August 1997. American Statistical Association.

[38] S. D. Conte, H. E. Dunsmore, and V. Y. Shen. *Software Engineering Metrics and Models*. The Benjamin/Cummings Publishing Company Inc., Menlo Park CA, 1986.

[39] P. B. Crosby. *Quality is Free: the Art of Making Quality Certain*. New American Library, NY, 1980.

[40] M. K. Daskalantonakis. A practical view of software measurement and implementation experiences within Motorola. *IEEE Transactions on Software Eng.*, 18(11):998–1010, November 1992.

[41] G. DeJong and R. Mooney. Explanation-based learning: An alternative view. *Machine Learning*, 1(2):145–176, 1986.

[42] Tom DeMarco. *Why Does Software Cost So Much ?*, chapter 2: Mad About Measurement, pages 11–25. Dorset Housing Publishing, 1995.

[43] R. A. DeMillo and R. J. Lipton. Software project forecasting. In A. J. Perlis, F. G. Sayard, and M. Shaw, editors, *Software Metrics*. MIT Press, Cambridge MA, 1981.

[44] W. E. Deming. *Out of The Crisis*. MIT Center for Advanced Engineering Study. MIT Press, Cambridge MA, 1986.

[45] R. Dion. Process improvement and the corporate balance sheet. *IEEE Software*, 10(4):28–35, July 1993.

[46] K. El Eman, N. Moukheiber, and N. H. Madhavji. An empirical evaluation of the G/Q/M method. In *Proceedings of CASCON 93*, pages 265–289, Toronto ON, November 1993. IBM Canada Ltd.

[47] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34, November 1996.

[48] U. Fayyad and R Uthurusamy. Data mining and knowledge discovery in databases. *Communications of the ACM*, 39(11):24–26, November 1996.

[49] A. V. Feigenbaun. *Total Quality Control.* McGraw Hill, NY, 40th anniversary edition, 1991.

[50] N. E. Fenton and A. Melton. Deriving structurally based software measures. *J. Syst. Software*, 12(3):177–187, 1990.

[51] Norman E. Fenton. *Software Metrics: A Rigorous Approach.* Chapman Hall, 1991.

[52] Norman E. Fenton. Software measurement: A necessary scientific basis. *IEEE Transactions on Software Eng.*, 20(3), March 1994.

[53] L. Finkelstein and M. S. Leaning. A review of the fundamental concepts of measurement. *Measurement*, 2(1), January 1984.

[54] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, pages 57–70, Fall 1992.

[55] T. Gilb. *Principles of Software Engineering Management.* Addison Wesley, 1987.

[56] C. Glymour, D. Madigan, D. Pregibon, and P. Smyth. Statistical inference and data mining. *Communications of the ACM*, 39(11):35–41, November 1996.

[57] P. Goodman. The practical implementation of process improvement initiatives. In N. Fenton, R. Whitty, and Y. Iizuka, editors, *Software Quality Assurance and Measurement: A Worldwide Perspective.* International Thomson Computer Press, 1995.

[58] R. B. Grady. *Practical Software Metrics for Project Management and Process Improvement*, chapter 3. Hewlett-Packard Professional Books, 1992.

[59] R. B. Grady. *Practical Software Metrics for Project Management and Process Improvement*, chapter 1. Hewlett-Packard Professional Books, 1992.

[60] Robert B. Grady. Successfully applying software metrics. *IEEE Computer*, 27(9):18–25, September 1994.

[61] Tracy Hall and Norman Fenton. Implementing effective software metrics programs. *IEEE Software*, 14(2):55–65, March 1997.

[62] W. Harrison. Software measurement: A decision support approach. In *Advances in Computers*, volume 39, pages 51–105. Academic Press Inc., 1994.

[63] J. Hersleb, D. Zubrow, D. Goldenson, W. Hayes, and M. Paulk. Quality improvement and the capability maturity model. *Communications of the ACM*, 40(6):31–40, June 1997.

[64] J. H. Holland, K. J. Holyoak, R. E. Nisbett, and P. R. Thagard. *Induction: Processes of Inference, Learning and Discovery.* MIT Press, Cambridge MA, 1986.

[65] M. Holsheimer and A. P. J. Siebes. Data mining: the search for knowledge in databases. Technical Report CS-R9406, CWI - Department of Algorithms and Architecture, Amsterdam, The Netherlands, 1994.

[66] W. S. Humphrey. *Managing the Software Process.* SEI Series in Software Engineering. Addison-Wesley, Reading MA, 1989.

[67] IBM Data Management Solutions. IBM's data mining technology. White Paper, 1996.

[68] Y. Iizuka. A new paradigm for software quality: the turning point for the japanese software. In N. Fenton, R. Whitty, and Y. Iizuka, editors, *Software Quality Assurance and Measurement: A Worldwide Perspective.* International Thomson Computer Press, 1995.

[69] J. M. Juran. *Juran on Planning for Quality.* Free Press, New York, 1988.

[70] Y. Kaneko, Y. Kadota, and S.Ohba. Behaviour analysis makes the company mature. In N. Fenton, R. Whitty, and Y. Iizuka, editors, *Software Quality Assurance and Measurement: A Worldwide Perspective.* International Thomson Computer Press, 1995.

[71] C. F. Kemerer. An empirical validation of software cost estimation models. *Communications of the ACM*, 30(5):416–429, May 1987.

[72] T. M. Khoshgoftaar, E. B. Allen, and D. L. Lanning. An information theory-based approach to quantify the contribution of a software metric. *Journal Systems and Software*, 36(2):103–113, February 1997.

[73] Barbara Kitchenham, Shari L. Pfleeger, and Norman E. Fenton. Towards a framework for software measurement validation. *IEEE Transactions on Software Eng.*, 21(12):929–944, 1995.

[74] Barbara Kitchenham, Lesley Pickard, and Shari L. Pfleeger. Case studies for method and tool evaluation. *IEEE Software*, 12(4):52–62, July 1995.

[75] W. Klösgen and J. M. Żytkow. Knowledge discovery in databases terminology. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, Cambridge, MA, 1996.

[76] M. Kogure and Y. Akao. Quality function deployment and CWQC in Japan. *Quality Progress*, pages 25–29, October 1983.

[77] R. Likert. *The Human Organization: Its Management and Value*. McGraw Hill, NY, 1967.

[78] Y. S. Lincoln and E. G. Guba. *Naturalistic Inquiry*. Sage, 1985.

[79] R. P. Lippmann. An introduction to computing with neural nets. *IEEE ASSP Magazine*, pages 4–22, April 1987.

[80] T. J. McCabe. A complexity measure. *IEEE Transactions on Software Eng.*, 2(4):308–320, April 1976.

[81] J. A. McCall, P. K. Richards, and G. F. Walters. Factors in software quality. Technical Report TR-77369, RADC, 1977.

[82] F. McGarry. Top-down x bottom-up process improvement. *IEEE Software*, 11(4), April 1994.

[83] A. Melton, D. Gustafson, J. Bieman, and A. Baker. A mathematical perspective for software measures research. *J. of Software Eng.*, 5(5):246–254, 1990.

[84] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.

[85] T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine Learning*, 1(1):47–80, 1986.

[86] B. Muller and J. Reinhardt. *Neural Networks, An Introduction*. Springer-Verlag, Berlin, 1991.

[87] S. Murthy, S. Kasif, S. Salzberg, and R. Beigel. OC1: Randomized induction of oblique decision trees. In *Proceedings of the 11th National Conference on Artificial Intelligence*, pages 322–327, July 1993.

[88] Raymond J. Offen and Ross Jeffery. Establishing software measurement programs. *IEEE Software*, 14(2):45–53, March 1997.

[89] Paul Oman and Shari L. Pfleeger. *Applying Software Metrics*. IEEE Computer Society Press, Los Alamitos CA, 1997.

[90] M. C. Paulk, B. Curtis, M. B. Chrissis, and C. V. Weber. Capability maturity model, version 1.1. *IEEE Software*, 10(4), July 1993.

[91] Shari L. Pfleeger. Lessons learned in building a corporate metrics program. *IEEE Software*, 10(3):67–74, May 1993.

[92] H. Potier, J. Albin, R. Ferreol, and A. Bilodeau. Experiments with computer complexity and reliability. In *Proceedings of the 6th International Conference on Software Engineering*, pages 94–103, Tokyo, Japan, September 1982. IEEE Computer Society Press.

[93] R. S. Pressman. *Software Engineering: A Practitioner's Approach*, chapter 1. McGraw Hill Inc., 3rd edition, 1992.

[94] J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

[95] R. Radice, A. J. Harding, P. E. Munnis, and R. W. Phillips. A programming process study. *IBM Systems Journal*, 24(2), 1985.

[96] S. Rifkin and C. Cox. Measurement in practice. Technical Report CMU/SEI-91-TR-16, SEI, 1991.

[97] F. S. Roberts. *Measurement Theory with Applications to Decision Making, Utility, and the Social Sciences*, chapter 1. Addison Wesley Inc., 1979.

[98] R. J. Rubey and R. D. Hartwick. Quantitative measurement of program quality. In *Proceedings of the 23rd ACM National Conference*, pages 671–677, Princeton NJ, 1968. Brandon/Systems Press.

[99] J. C. Schilimmer and P. Langley. Machine learning. In S. C. Shapiro, editor, *Encyclopedia of Artificial Intelligence*, volume 1, pages 785–805. John Wiley & Sons, 1992.

[100] N. F. Schneidewind. Methodology for validating software metrics. *IEEE Transactions on Software Eng.*, 18(5):410–422, May 1992.

[101] R. Selby and A. H. Porter. Learning from examples: Generation and evaluation of decision trees for software resource analysis. *IEEE Transactions on Software Eng.*, 14(12):1743–1757, December 1988.

[102] M. Shepperd. Algebraic models and metric validation. In I. Somerville and M. Paul, editors, *Formal Aspects of Measurement*, Lecture Notes in Computer Science. Springer Verlag, 1992.

[103] W. A. Shewhart. *Economic Control of Quality of Manufactured Product.* D. Van Nostrand Company, Inc., New York, 1931.

[104] K. Srinivasan and D. Fisher. Machine learning approaches to estimating software development effort. *IEEE Transactions on Software Eng.*, 21(2):126–137, February 1995.

[105] S. S. Stevens. On the theory of scales of measurement. *Science*, 103:677–680, 1946.

[106] M. Thomas. Top-down x bottom-up process improvement. *IEEE Software*, 11(4), April 1994.

[107] J. Tian, J. Henshaw, and I. Burrows. Analysis of factors affecting in-field product quality using tree-based predictive modeling. In *Proc. IBM Software Development Conference*, San Jose, California, May 1994.

[108] J. Tian and M. Zelkowitz. A formal program complexity model and its application. *J. Syst. Software*, 17:253–266, 1992.

[109] C. E. Walston and C. P. Felix. A method of programming measurement and estimation. *IBM Systems Journal*, 16(1):54–73, January 1977.

[110] Edward F. Weller. Using metrics to manage software projects. *IEEE Computer*, 27(9):27–33, September 1994.

[111] E. J. Weyuker. Evaluating software complexity measures. *IEEE Transactions on Software Eng.*, 14(9):1357–1365, September 1988.

[112] J. P. Womack, D. T. Jones, and D. Roos. *The Machine That Changed the World: Based on the MIT 5-Million Dollars 5-year Study on the Future of the Automobile.* Rawson Associates, NY, 1990.

[113] Marvin V. Zelkowitz and Dolores Wallace. Experimental models for validating computer technology. *IEEE Computer*, to appear.

[114] H. Zuse. *Software Complexity: Measures and Methods.* deGruyter, 1990.