Validation of an Approach for Improving Existing Measurement Frameworks

Manoel G. Mendonça and Victor R. Basili, *Fellow*, *IEEE*

Abstract—Software organizations are in need of methods to understand, structure, and improve the data they are collecting. We have developed an approach for use when a large number of diverse metrics are already being collected by a software organization [1], [2]. The approach combines two methods. One looks at an organization's measurement framework in a top-down goal-oriented fashion and the other looks at it in a bottom-up data-driven fashion. The top-down method is based on a measurement paradigm called Goal-Question-Metric (GQM). The bottom-up method is based on a data mining technique called Attribute Focusing (AF). A case study was executed to validate this approach and to assess its usefulness in an 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. It identified several new metrics for the interviewed groups, and also contributed to better understanding the data user needs. The bottom-up method was used to gain new insights into the existing CUSTSAT data. Unexpected associations between key variables prompted new business insights, and revealed problems with the process used to collect and analyze the CUSTSAT data. This paper uses the case study and its results to qualitatively compare our approach against current ad hoc practices used to improve existing measurement frameworks.

Index Terms—Software metrics, goal-oriented measurement, GQM, data mining, knowledge discovery, AF, experimental validation, method evaluation, case study.

1 Introduction

There are many different groups involved in the processes of developing, maintaining, and managing software. Those groups need to use measurement to characterize, control, predict, and improve those processes. We define a Measurement Framework (MF) as a set of related metrics, data collection mechanisms, and data uses inside a software organization.

In general, software organizations have evolved their measurement frameworks over time, based upon input from a variety of sources and needs, without a well structured set of goals. This scenario can lead to poorly structured measurement and data use. Software organizations can 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 redundant data; 2) collecting data that nobody uses; or 3) collecting data that might be useful to people who do not even know it exists inside their organization. For these reasons, improving ongoing measurement is an important problem for many software organizations. We believe the solution for this problem needs to address two key issues: 1) to better understand and structure this ongoing measurement; 2) to better explore the data that the organization has already collected.

We proposed an approach that addresses these two critical issues jointly. 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 GQM-based method is used to understand and structure ongoing measurement, and an AF-based method is used to discover new interesting information in the legacy data.

We validated this approach through a case study in an industrial setting. We used our approach to analyze the customer satisfaction (CUSTSAT) survey data at the IBM Toronto Laboratory. This paper introduces, the approach and presents its validation through the case study we performed at IBM's customer satisfaction measurement framework (referred simply as CUSTSAT MF from now on). The paper is organized as follows: Section 2 discusses the related work and basic concepts on which the approach is based. Section 3 describes the approach itself. Section 4 presents the approach validation at the CUSTSAT MF. Section 5 contains the conclusions and final remarks.

2 Background and Related Work

Our work is based on the premise that a good measurement framework should be sound, complete, lean, and consistent. An MF is **sound** when its metrics and measurement models are valid in the environment where they are used. An MF is **complete** when it measures everything that its users need to achieve their goals. An MF is **lean** when it measures what is needed and nothing else (metrics cost money to collect [3]). An MF is **consistent** when its metrics are consistent with the

- *M.G. Mendonça is with the Computer Networks Research Group, Salvador University (UNIFACS), Avenue Cardeal da Silva 747, Salvador, Ba, Brazil, 40220-141. E-mail: mgmn@unifacs.br.*
- *V.R. Basili is with Fraunhofer Center for Experimental Software Engineering and the Department of Computer Science, University of Maryland, College Park, MD 20742. E-mail: basili@cs.umd.edu.*

Manuscript received 4 Aug. 1998; accepted 25 Nov. 1999.
Recommended for acceptance by D.R. Jeffery.
For information on obtaining reprints of this article, please send e-mail to: tse@computer.org, and reference IEEECS Log Number 107232.

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, leanness, and consistency of measurement frameworks is not a new idea in software measurement. In a seminal 1976 work, Boehm, et al. [4], wrote:

> "Our . . . approach were as follow: 1. Determine a set of characteristics which are important . . . and reasonably exhaustive and nonoverlapping . . . 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 has been concerned with the issue of using sound metrics.

Metrics have been validated in very different ways. Analytical validation has been used to: 1) analyze if a metric is theoretically sound [5], [6], [7], [8]; or 2) verify if a metric fulfills the properties that are associated with the attribute it is supposed to measure [9], [10], [11], [12], [13]. Empirical validation of predictive models has been used to validate these models' precision and accuracy [14], [15], [16], [17]. Empirical validation of direct metrics has been used to: 1) analyze the association between these metrics and important quality measures [18], [19], [20], [21], [22]; and 2) assess these metrics consistency when they are used by different people to measure the same thing [22], [23].

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 [24], [25], [26], [27], [28]. Only recently, methodologies have been proposed to build complete, lean, and consistent MFs [29], [30]. 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 paper deals precisely with these issues.

2.1 Terminology and Basic Concepts

We will adopt a consistent terminology throughout this paper. This terminology is derived from the data mining terminology proposed by Klösgen and Zytkow [31] and the software engineering measurement terminology proposed by Fenton [32]. During this section (and the rest of this paper), **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 an 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. 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 paper, a goal is always described using the template we will introduce in Section 2.2. **Domain knowledge** is nontrivial 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. **New** or **discovered knowledge** is the new domain knowledge that data users gain by analyzing the data.

2.2 The GQM Paradigm

The Goal-Question-Metric Paradigm was proposed as a means of measuring software in a purposeful way [33], [34]. The GQM paradigm first step is to 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., Motorola [24], NASA [35], HP [36], AT&T [37]).

Fig. 1 shows an abstract example of what we call a GQM structure. The following template, defined by Basili and Rombach [33], is used to define measurement goals:

$$\begin{aligned}&\text{Analyze } '\underline{\text{object of study}}' \text{ in order to } '\underline{\text{purpose}}'\\ &\text{with respect to } '\underline{\text{focus}}' \text{ from the point of view}\\ &\text{of } '\underline{\text{point of view}}.'\end{aligned} \tag{1}$$

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

$$\begin{aligned}&\text{Analyze } '\text{service support for our product}'\\ &\text{in order to } '\text{evaluate it}' \text{ with respect to}\\ &'\text{customer satisfaction}' \text{ from the point}\\ &\text{of view of } '\text{service support personnel}.'\end{aligned} \tag{2}$$

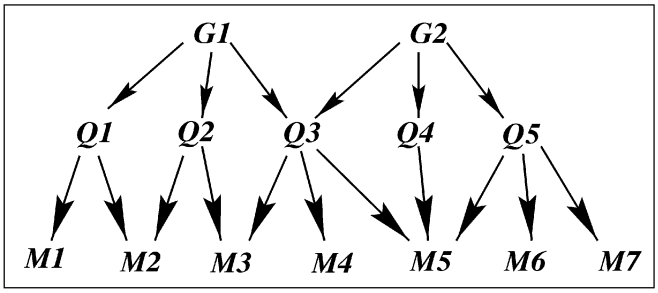

Fig. 1. An abstract GQM structure.

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.

The GQM is a general paradigm that has been instantiated in several different ways [35], [36], [37], [38], [39]. All those instantiations aim to define measurement from scratch. This paper 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.

In our approach, each GQM structure will specify the goals associated with a certain data user group (goals with the same "point of view"). Each structure will allow us to trace the goals of a certain user group to the measures that are intended to define them operationally. It will also provide a platform to interpret the data and better understand the data user needs.

2.3 The Attribute Focusing Technique

Attribute Focusing (AF) is a data mining technique that has been used in several different applications including: software process [40], [41], [42], customer satisfaction [43], and sports [44] 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.

Fig. 2 shows an example of an Attribute Focusing diagram. It was obtained from a real data set pertaining to a particular class of software products [43]. 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 approximately 56.5 percent of those "involved in the decision to buy the product" were "very satisfied with the product."

The diagram in Fig. 2 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.

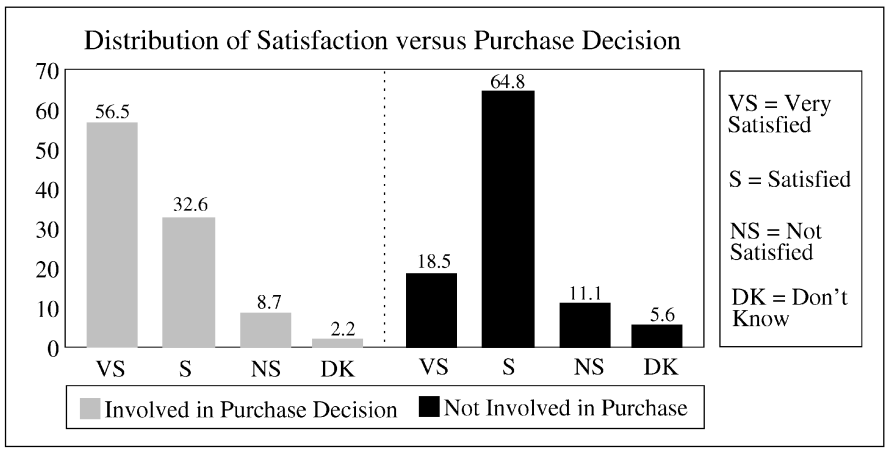

Fig. 2. A two-way attribute focusing diagram.

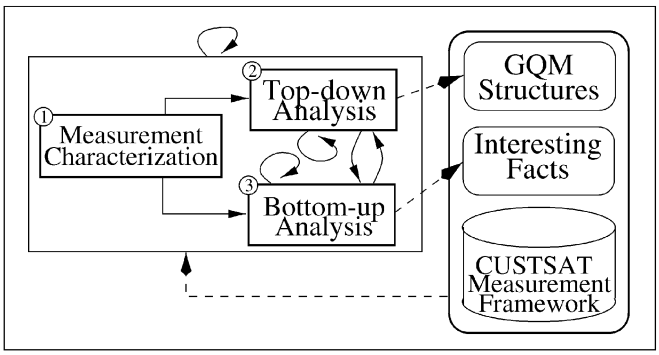

Fig. 3. The approach.

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.

2.3.1 The Interestingness Function

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

$$Interestingness(A_x, A_y) = \forall u \forall v \{\max[In_2(A_x = v; A_y = u)]\}. \quad (3)$$

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

$$(Observed(A_x = v \wedge A_y = u)) :$$

$$In_2(A_x = v; A_y = u) = |Observed(A_x = v) \times Observed(A_y = u) - Observed(A_x = v \wedge A_y = u)|. \quad (4)$$

$Observed(A_x = v)$ 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. The formalism for two-way relationships can be extended to N-way relationships. Consider three-way relationships as an example:

$$In_3(A_x = v; A_y = u; A_z = t) = |Obs(A_y = u \wedge A_z = t) \times Obs(A_x = v) - Obs(A_x = v \wedge A_y = u \wedge A_z = t)|. \quad (5)$$

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 two-way relationships between A_x, A_y, and A_z. In other words:

$$Interestingness(A_x, A_y, A_z) > Interestingness(A_x, A_y) \text{ and } Interestingness(A_x, A_y, A_z) > Interestingness(A_x, A_z) \quad (6)$$

In our work, we have used functions that estimate the interestingness level for associations among up to four attributes. For further discussion on AF analyses between an arbitrary number of attributes (N-way analyses) and the concept of "Interestingness," the interested reader should consult [2] and [45].

3 The Approach and Its Application

As mentioned before, the purposes of our approach are to better: 1) understand the ongoing measurement, 2) structure it, and 3) explore the MF legacy data. For that, our approach is divided into three different phases, namely: measurement framework (MF) characterization, top-down analysis, and bottom-up analysis. The approach is depicted in Fig. 3.

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 MF. The third phase, bottom-up analysis, is based on the AF technique. It is executed to extract knowledge (useful, interesting, and nontrivial information) from the already existing data.

Fig. 3 shows the information flow (dashed lines) and control flow (solid lines) of this process. The two main products of our approach are: 1) GQM structures, produced by the top-down analyses; and 2) interesting facts, produced by the bottom-up analyses.

The control flow, described by solid arrows in Fig. 3, is determined by the interaction 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 prerequisite 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.

The top-down and bottom-up analyses are designed to be applied incrementally. Our basic unit of analysis is a user group (also called a point of view). This makes it possible to use our approach to incrementally improve large MFs, one point of view at a time.

The approach described in Fig. 3 was applied to improve the IBM Toronto Laboratory's Customer Satisfaction (CUSTSAT) measurement framework. In the CUSTSAT MF, 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 Laboratory, the CUSTSAT data is used by several different groups (e.g., development, service, support, and senior management). The large amount of data and the diversity of

groups that are interested in this data made the CUSTSAT MF a good subject for applying our improvement approach.

The sections that follow describe the processes applied during the three phases of our approach. Those processes are described in greater detail in [1]. The readers interested in a complete description and discussion of those processes should refer to [2, Chapter 3].

3.1 The Measurement Framework Characterization

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

We used a combination of structured interviews [46] and review of the available MF documents to capture and document those key components at CUSTSAT MF. We executed the following process to characterize the CUSTSAT measurement framework:

Step 1: Identifying Metrics and Attributes. The first components identified were the metrics used the CUSTSAT MF and the attributes they were trying to measure. Those tasks were simple. Most of the metrics corresponded to questions in the survey questionnaire. The metrics' meanings, corresponding to the attributes we believed the metrics were measuring, could be identified in the formulation of the questions in the survey questionnaire. Terms like "capability," "performance," or "maintainability" were explained when they were used.

Step 2: Identifying Available Data. The second type of component to be identified was the data available in the MF. We listed when and under what circumstances the metrics were used to collect customer satisfaction data, where the resulting data was stored, and how to access it.

Step 3: Identifying Data Uses and User Groups. The third type of components to be identified are the data uses and data user groups. We interviewed the CUSTSAT data manager to do that. We used a checklist for the information we wanted to collect during the interviews. We started by:

1. Listing the data analyses and data presentations that used the CUSTSAT data; and
2. Identifying the people who used those analyses or were present at those data presentations.

Each type of data analysis or presentation (DA/P) was described as a distinct data use. The data use descriptions included the frequency with which the DA/Ps were done, the list of metrics used in them, the granularity and scope of the DA/Ps, and the list of groups that took part in the DA/Ps.

The 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 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 them.

3.2 The 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 to gain better understanding of the data user needs. The top-down analysis uses a method based on the Goal-Question-Metric Paradigm (see Section 2.2). This GQM-based method is applied to build (or revise) a structure that maps the goals of a data user group to the metrics (and data) used in the organization. This structure is used to identify missing or extraneous elements of a MF from the user group's point of view. We interviewed representatives of the data user groups to build such GQM structures for them.

In the CUSTSAT MF, we applied our GQM-based method to a limited number of data user groups as our main objective was 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 groups chosen are associated with the database product development at the laboratory:

1. The DB customer service and support group.
2. The DB usability (user interface design) group.
3. The DB information development (documentation) group.

In this paper, we will describe the building of the GQM structure for the DB service support group to illustrate how we have applied the GQM-based method in the CUSTSAT MF. This group gives vendor support to the client's database installations. Its responsibilities are to give fast resolution for client problems and provide permanent solutions to prevent these problems from recurring.

We used a structured interview [46] to build the GQM structure for the service support group. We interviewed a senior representative of the group. All the material for the interviews were prepared beforehand. It included:

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

Step 1: Capturing the User Group Goals. We used the goal template described in Section 2.2 to capture the goals of the user group. For each goal, we had to identify the goal's "object of study," "purpose," and "focus" (the "point of view" is the user group itself). The first part of this step was to discuss the data analyses and presentations (DA/Ps) done for the group. This allowed us to:

1. Motivate and focus the rest of the interview around the CUSTSAT MF; and

2. Validate our understanding of their data usage (including assessing the importance of the data for them).

Next, we captured their goals in using the CUSTSAT data. We asked the group representative 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 customer in order to understand them with respect to expectations with support service.

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

Step 2: Identifying Relevant Entities and Attributes. The next step was to identify the entities and attributes the user group wanted to measure to achieve their goals: what we call "relevant entities" and "relevant attributes." We started by identifying the relevant entities. 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 identified other relevant entities by finding out which entities are related to the "object of analysis" and which may affect the "goal focus" from the data user group point of view.

For each relevant entity, we prepared an initial list of attributes that might be relevant for the stated goal. In order to produce a comprehensive list of attributes for each entity, we used a checklist based on the entity type. The initial list of relevant attributes was then reviewed and expanded by the user group representative during the interviews. The end result was a list of attributes classified according to their relevance to the user group's goals.

Step 3: Mapping Attributes to Existing Metrics. The last step was to map the relevant attributes to metrics that were being used in the organization. Remember that an attribute states "what" we want to measure while the metrics defines "how" we measure something. The mapping consisted of checking if the metrics were measuring the things (attributes) the user group wants to measure.

At this step, a GQM structure was assembled for the user group. Fig. 4 depicts the GQM structure for the service support group. This structure shows the mapping between the user goals, the relevant entities, the relevant attributes, and the metrics used in the MF. It documents the user group's needs measurement-wise. 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 Fig. 4, 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. The crossed out metrics: Q45c, Q45a, Q45b, and Q6a indicate that their associated attributes are not relevant to the service support group.[1] They are extraneous from the service support point of view. At the end of this step, we had a list of inconsistent, missing, and extraneous metrics from the user group's point of view.

3.3 The Bottom-Up Analysis

Bottom-up analyses are aimed at discovering new and useful information in the existing data, thus, improving data awareness and data usage. The key feature of bottom-up analyses is a shift from hypothesis driven data analysis to discovery driven data analysis. Traditionally, the goal of extracting information from data has been achieved by combining hypothesis formulation and data collection. Under this schema, a domain expert must hypothesize the existence of information of interest, gather data to test this hypothesis, analyze the data, and interpret the results. The last two steps are usually done with statistical data analysis techniques.

Due to the complexity and amount of data stored in a large measurement framework, the hypothesis driven approaches are usually not sufficient to fully explore the information contained in the MF's data. Hypothesis driven approaches should be combined with discovery driven approaches. Those approaches have the ability to automatically discover important information hidden in the data and present it in an appropriate way to be interpreted by a domain expert.

The idea of using discovery driven data analysis is not new to our field. The literature has many examples of the use of machine learning techniques to extract knowledge (new and useful information) directly from software engineering data sets [15], [47], [48], [49], [50], [51]. In our case, the bottom-up analyses use a method based on a data mining technique, called Attribute Focusing [52], to extract unexpected and useful information directly from the MF database. This "AF-based method" establishes procedures to effectively apply the AF technique, maximizing knowledge discovery, and minimizing discovery cost.

Section 2.3 introduced the AF technique. The technique produces a set of ordered interesting diagrams to be examined by "experts" in a given knowledge domain. In the case of a measurement framework, those "experts" correspond to the MF data users. In this context, the AF-based method allows the data users 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 three steps. In the first step, the people in charge of applying the bottom-up method to the legacy data (i.e., data analysts) interact with the data users

1. The service support group wanted to understand the customer expectations with the support services (Goal 2), but did not consider the customer organization and contact person relevant to this goal.

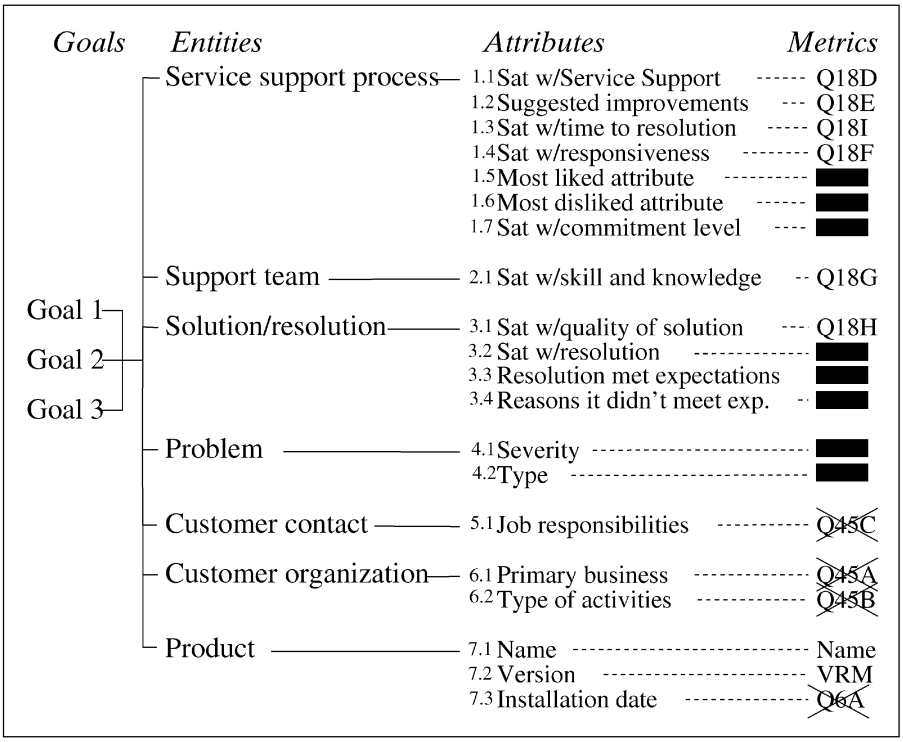

Fig. 4. GQM structure for the Service Support Group.

to define the type of analysis that will be done. In the next step, 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 is when knowledge discovery takes place.

Step 1: Defining the Analysis. AF investigates several empirical relations in each analysis. We use a Generic Relationship Question (GRQ) to state the set of relations to be investigated empirically avoiding the computation of uninteresting relations. The following template is used to define a GRQ:

$$How\ do\ 'Attribute\ class\ X_1'\ and\ \ldots\ and\ 'Attr.\ class\ X_{N-1}'\ [relate\ to,\ affect,\ impact]\ 'Attribute\ class\ Y'?' \tag{7}$$

An attribute class defines set of attributes grouped according to certain criteria or features relevant to a user group. In each analysis, only the empirical relations between attributes from different product classes are investigated. Consider the following GRQ, defined for an AF analysis done at the CUSTSAT measurement framework.

$$How\ does\ the\ ''local\ support\ and\ product\ features\ satisfaction''\ relate\ to\ the\ ''most\ important\ attributes?'' \tag{8}$$

In the above GRQ, local support (LSSATs) and product features (PFSATs) satisfaction are grouped in class "X_1," while the most important satisfaction attributes (MIAs) are grouped in class "Y." The LSSATs group contains attributes such as: satisfaction with training, local sales, and local technical support. The PFSATs group contains attributes such as: satisfaction with the product's performance, capability, reliability, and usability. The MIAs group contains the attributes that senior management consider most important for IBM. It includes attributes such as: overall satisfaction with product (OSAT), whether the customer would recommend the product to someone else, whether the customer is planning to upgrade the product, etc. We used the above GRQ to define a comparison between local support and product features attributes with respect to their impact on the most important attributes in the CUSTSAT MF.

After establishing a GRQ for an AF analysis, the analysis itself must be defined. First, the attributes identified in a GRQ have to be mapped to the metrics in the MF. In our example, the above GRQ was used to identify the questions (metrics) of interest in the survey questionnaire. Second, the data granularity and scope of analysis has to be derived from the user group goals and/or data use descriptions. The data sets can be extracted and formatted for analysis after that. In our example, the granularity and scope of the analysis defined a data set containing all data points collected in 1995 for database products (IBM and competition).

Step 2: Running the Analysis and Organizing the Diagrams. The next step is to run the AF tool to produce diagrams to the data users. This step is almost completely automated. The inputs are:

1. metric groupings;
2. maximum number of diagrams (relations) to be produced; and
3. analysis dimension.

TABLE 1
LSSAT and PFSAT Impact on the Most Important Attributes

MIA	LSSAT and PFSAT impact on the MIA			
	Positive Impact (Very satisfied with PFSAT or LSSAT ⇒ New VS)		**Negative Impact** (Not satisfied with PFSAT or LSSAT ⇒ New NS)	
		New VS		**New NS**
Original VS 33.2% Attribute X **Original NS** 8.6%	PFsat1	67.2%	Rsat	42.6%
	PFsat2	65.6%	PFsat1	41.4%
	PFsat3	61.6%	PFsat2	29.3%
	PFsat4	59.0%	PFsat3	29.3%
	Rsat	57.6%	LSsat1	20.8%
	LSsales	51.9%	PFsat4	20.2%
	LSsat1	51.8%	LSsales	15.4%
	LSsat2	50.3%	LSsat2	14.3%
Original VS 48.7% Attribute Y **Original NS** 13.3%	PFsat1	72.1%	PFsat1	44.9%
	LSsales	69.4%	Rsat	38.0%
	PFsat2	68.9%	PFsat3	37.7%
	Rsat	66.6%	PFsat2	35.0%
	PFsat3	66.4%	LSsat1	28.6%
	LSsat1	65.6%	PFsat4	27.8%
	LSsat2	64.5%	LSsales	21.8%
	PFsat4	61.1%	LSsat2	16.8%

The groupings are directly derived from the attribute classes defined by the GRQ. 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 three-way diagrams). After all parameters are entered, the analysis can finally be run and, usually, a sizable set of diagrams is produced.

Although many uninteresting diagrams are pruned away with the metric groupings (defined by the attribute classes), 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 affecting the interestingness values and driving the discoveries.

After a sizable number of useful diagrams have been compiled, we organize them to facilitate the data user's inspection. We can group diagrams according to several criteria (see [2, Chapter 3], for details). This procedure produces several "groups of diagrams" to be shown to the data users.

In our example analysis, we grouped diagrams according to the positive and negative impacts of LSSATs and PFSATs 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 PFSAT or a LSSAT 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 PFSAT or a LSSAT attribute.

Table 1 shows the summary of positive and negative impacts of the LSSATs and PFSATs in two particular MIAs, Attribute X and Y. For example, the first line of Table 1 shows that 67.2 percent of the customers who were very satisfied, with respect to PFSAT1, were also very satisfied with Attribute X. On the same token, 42.6 percent of the customers who were not satisfied, with respect to RSAT, were also not satisfied with Attribute X. Table 1 shows two attributes explicitly: RSAT (satisfaction with product reliability, a PFSAT), and LSsales (satisfaction with local sales support, a LSSAT). The others PFSATs (PFSAT1, PFSAT2, PFSAT3, and PFSAT4) and LSSATs (LSSAT1 and LSSAT2) are not made explicit to protect IBM proprietary information.

Step 3: Reviewing the Diagrams. The last step of the AF-based method is the analysis of the diagram groups by the data users. The diagram groups have many types of information in them:

1. Unexpected correlations between metrics (direct analysis of N-way diagrams).
2. Unexpected value distributions (direct analysis of one-way diagrams).
3. Unexpected (in)consistencies in the relationships between explanatory metrics and related explained metrics (direct from analysis of a diagram group).

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.

In our example, the analysis produced several interesting results:

- For some MIAs, LSSATs are sometimes as important as PFSATs like product performance or reliability satisfaction. For example, Table 1 shows that "local sales support" (a LSSAT) has a higher positive impact than "reliability" (a PFSAT) with respect to "Attribute Y."
- The same PFSATs and LSSATs had different types of impacts in different MIAs. For example, "local sales support" (a LSSAT) was one of the attribute with the highest positive association with "Attribute Y" (a MIA), while it was one of the attributes with the lowest positive associations with "Attribute X" (another MIA). This was a surprise because there used to be an implicit assumption

that the PFSATs and LSSATs 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 "Attribute X."

At the CUSTSAT MF, 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.

4 Work Validation

There are several experimental methodologies to validate new software technologies [53]. However, we believed that the nature of our technology, an approach useful to improve large measurement frameworks, required that it be applied as a case study in an industrial setting [54]. The chosen environment, the CUSTSAT MF at the IBM Toronto Laboratory, was one in which a large measurement framework existed and was being used. This allowed us to compare the results of our approach with the existing process used to improve the CUSTSAT MF.

During the case study, we used a combination of evaluation procedures to compare our approach with the processes that were already being used to improve the CUSTSAT MF. We combined quantitative and qualitative analyses to evaluate the effect of the approach's methods. We used "qualitative effects analysis" [55] to assess quantitatively and qualitatively the effect of the methods according to the subjective opinion of the CUSTSAT data manager. We also used some direct measures to quantitatively validate the data manager's expert opinion.

4.1 Objectives of Our Approach and Validation Goals

Our work addressed three key issues: 1) better understanding the ongoing measurement; 2) better structuring it; and 3) better exploring the data that the organization has already collected. It did not intend to be a comprehensive or definitive approach to improve measurement frameworks. Our work objectives were:

- **O**1 Discovering interesting data distributions and associations in the MF database.
- **O**2 Visualizing data distributions and associations in the MF database.
- **O**3 Assessing the importance of metrics for specific user groups and for the organization as a whole.
- **O**4 Assessing the structure (i.e., measurement instrument, scale, and domain value) of metrics used in the MF.
- **O**5 Assessing the appropriateness of the data collection process.
- **O**6 Assessing the importance of data analyses for specific user groups and for the organization as a whole.
- **O**7 Understanding and documenting the needs of users with respect to existing metrics, data analyses, and data presentations.
- **O**8 Understanding and documenting the measurement goals of the MF data users.
- **O**9 Identifying new applications and user groups for the data.
- **O**10 Identifying the need for new metrics, data analyses, and data presentations.

We did not expect our approach to completely fulfill all these objectives. The case study aimed 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 in the case study; and 3) evaluate the cost at which our approach fulfilled those objectives in the case study. Based on these goals, a set of objective and subjective validations was defined:

- **V**1 In order to achieve the first goal (relevance of the objectives), the IBM CUSTSAT data manager was asked to subjectively judge how important each of the listed objectives is to improving the CUSTSAT measurement framework.
- **V**2 In order to achieve the second validation goal (approach effectiveness), we:
  - **V**2a 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 MF's existing improvement process.
  - **V**2b Compared the direct impact of the use of the approach on the CUSTSAT measurement framework with its existing improvement process.
- **V**3 In order to achieve the third validation goal (the approach cost), we:
  - **V**3a Asked the data manager to subjectively judge how cost effective the three steps of the approach were.
  - **V**3b Measured how much effort was needed to apply the steps that compose our approach, and compared it with the effort to apply the existing improvement process.

V1 is referred to as the validation of the objectives relevance, **V**2a and **V**2b are referred to as the validation of the approach effectiveness, and **V**3a and **V**3b are referred to as validation of the approach cost effectiveness. **V**1, **V**2a, and **V**3a were based on subjective evaluations. **V**2b and **V**3b were based on objective evaluations.

4.2 Validation Questionnaire and Importance of the Improvement Objectives

The data for validations **V**1, **V**2a, and **V**3a was collected jointly through one questionnaire submitted to the CUSTSAT data manager at the IBM Toronto Lab. This questionnaire can be found in the appendix of reference [2].

The validation questionnaire had five point ordinal scale questions to evaluate the importance of our improvement objectives. These questions used the numbers 0-4 to quantify the improvement objectives. Number zero (0)

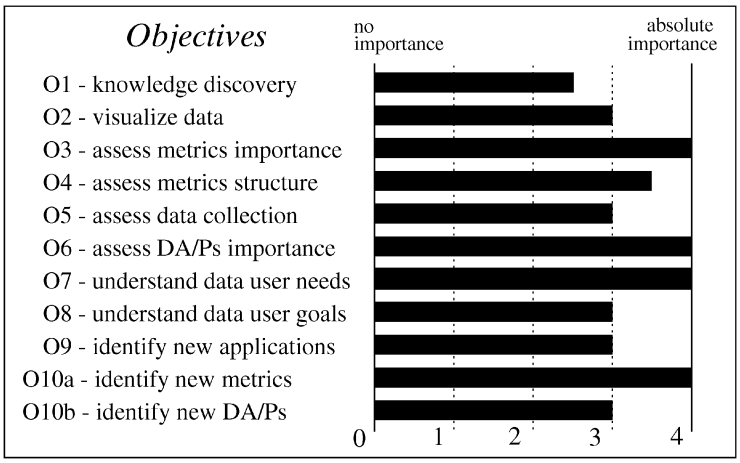

Fig. 5. Subjective rating of the improvement objectives.

meaning that the improvement objective has no importance at all. Number four (4) meaning that the improvement objective has absolute importance.

Fig. 5 summarizes the importance scores given by the data manager to our improvement objectives. It shows that, according to the data manager's subjective opinion, all the improvement objectives listed before are very relevant to the CUSTSAT MF.

4.3 Subjective Validation of Effectiveness

The subjective evaluation of the method's effectiveness was also based on the validation questionnaire. Quantitative and qualitative questions were used to subjectively compare the new improvement approach against the improvement process being used in the CUSTSAT MF. Like before, the quantitative questions used a 0-4 scale. The qualitative questions were open ended and asked for the data manager comments on the ratings he gave in the quantitative questions. The aim of these questions was to *qualitatively* determine what our approach added to the existing 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 existing process and the new improvement approach capabilities. The questions were not formulated for this purpose and one case study is not enough to make this type of comparison.

Table 2 summarizes the results from the subjective interview. Each row corresponds to one of the new approach improvement objectives. The first 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 existing process capabilities, the five point scores given by the data manager to the existing process was transformed in a three point scale (weak, some, and good capabilities).

Table 2 shows some important facts. The first one is that the CUSTSAT MF is mature. It has capabilities in several of the areas that the new approach proposes to improve. Nonetheless, we concluded that the capabilities that were already present in the MF were not the same as the ones provided by the new approach. The new approach complements or expands the MF capabilities even in areas where the MF already has good mechanisms helping to achieve the improvement objectives. Let us consider objective O10(b) (identifying the need for new DA/Ps) as an example. According to the data manager, the MF uses the channels that are open between the data managers and the data users (e.g., periodical data presentations) as mechanisms to successfully identify new applications for the existing data. However, new DA/Ps proposed by the data users usually emulate what is already done by other user groups inside the organization or try to further explore recognized hot areas of analysis. The AF-based method aims at discovering completely new areas to be explored by the organization. In this sense, its capabilities are complementary to the ones that already exist in the MF. For this reason, the data manager considered the AF-based method very helpful to identifying new DA/Ps (**O**10(b)) in the CUSTSAT MF.

TABLE 2
Summary of the Subjective Effectiveness Evaluation

Objective	MC	GQM	AF	Existing
O1	0	0	3.5	Weak mechanisms
O2	0	0	3	Weak mechanisms
O3(a) – user groups	3	4	0	Weak mechanisms
O3(b) – overall	0	2	3	Weak mechanisms
O4	0	0	1	Some mechanisms
O5	0	0	0	Some mechanisms
O6(a) – user groups	0	3	0	Good mechanisms
O6(b) – overall	0	3	0	Good mechanisms
O7	1.5	3	0	Some mechanisms
O8	0	3	0	Weak mechanisms
O9	0	0	3	Good mechanisms
O10(a) – new metrics	0	3	2	Good mechanisms
O10(b) – new DA/Ps	0	0	4	Good mechanisms

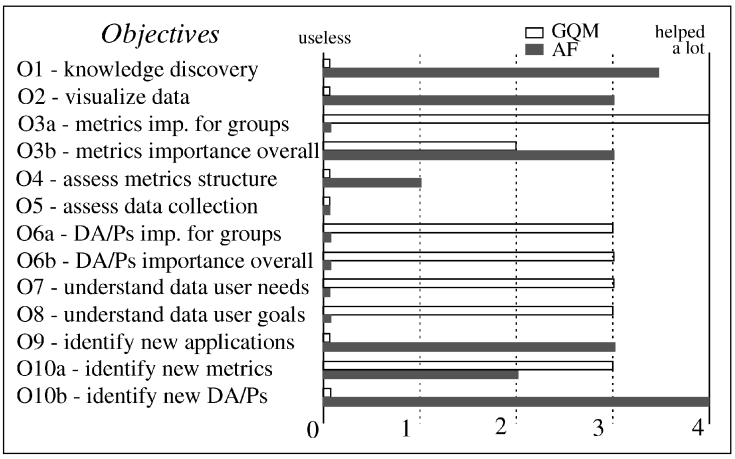

Fig. 6. A comparison between the AF and GQM-based methods.

Table 2 also shows that the new approach "failed" to meaningfully achieve objectives **O**4—assessing the structure of the metrics, and **O**5—assessing the structure of the questionnaire. According to the data manager, the new approach did little to help finding problems with the structure of the CUSTSAT questions and questionnaire. Another 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. 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 (**O**1); visualizing the data (**O**2); assessing the importance of metrics for specific user groups and the organization as a whole (**O**3); and understanding and documenting the data user goals (**O**8). The new approach significantly helped to achieve those objectives. The data manager considered that the AF-based method helped significantly to achieve objectives **O**1, **O**2, and **O**3(b). He also considered that the GQM-based method helped significantly to achieve objectives **O**3(a) and **O**8.

In summary, according to the data manager, 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 **O**3(a), **O**6(a), **O**6(b), **O**7, **O**8, and **O**10(a). The AF-based method helped significantly to achieve objectives **O**1, **O**2, **O**3(b), **O**9, and **O**10(b). This happens 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. Fig. 6 shows how the methods contributed to improve the measurement framework in several relevant and complementary aspects.

4.4 Objective Validation of Effectiveness

In order to validate the methods objectively, the impact of the new approach on the CUSTSAT survey questionnaire and on other parts of the measurement framework was analyzed. We considered the following factors: 1) Impact on the CUSTSAT questionnaire (relates to objectives **O**3, **O**4, **O**5, and **O**10); 2) impact on data usage (relates to objectives **O**6, **O**9, and **O**10); and 3) insights gained from data analysis (relates to objectives **O**1 and **O**2).

These factors were evaluated for each of the three phases of the new approach, and compared to the existing 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 also mostly qualitative.

4.4.1 Impact on the CUSTSAT Questionnaire

The AF analyses helped little to effectively modify the questionnaire. The GQM interview on the other hand produced the results summarized in Table 3. The table compares the modifications made because of ad hoc

TABLE 3
Impact of the GQM Interviews on the Questionnaire

Method	Year	Questions	modified	deleted	new
Service Support (SS) Group					
Existing	1995	6	0	0	4
Existing	1996	10	0	0	3
GQM	1996	10	0	0	8
Information Development (ID) Group					
Existing	1995	18	1	1	1
Existing	1996	18	1	1	2
GQM	1997	19	1	2	4
Usability (UI) Group					
Existing	1995	17	0	1	0
Existing	1996	16	0	0	0
GQM	1997	16	0	0	8

requests (in 1995 and 1996) against the modifications suggested by the GQM interviews. Consider the first row as an example. This row indicates that the 1995 questionnaire had six questions relevant to the service support group and that none of those were modified or deleted based on ad hoc requests. It also shows that four new questions were added to the questionnaire based on the 1995 ad hoc requests.

It is important to highlight that the table is comparing modifications made based on ad hoc requests by *all* groups that use the questionnaire, against modifications suggested by the GQM interviews with *particular* groups at the Toronto Laboratory.

The GQM interview suggestions have to be approved by an IBM division wide committee before they can be implemented. The interview with the UI and ID groups were conducted in 1997 and when this paper was written their suggestions were still subject to approval. The interview with the SS group was conducted in 1996. It suggested eight new questions (third row of Table 3), four of which were approved and implemented in 1997 questionnaire.

In order to evaluate the impacts of the GQM-based method in the questionnaire, let us discuss the context and meaning of these impacts. The service support group made four ad hoc metric requests in 1995. This was the main reason we decided to interview them in 1996. 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 1996 GQM interview with the Toronto SS group effectively produced four new metrics for 1997. It missed, however, three questions related to Internet service support that were later requested by a SS group from another laboratory.

Like the SS group, the ID group was very active in using the CUSTSAT data. However, their question set was 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 four new questions and drop two existing questions from the questionnaire. The suggestion to drop two 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. This indicated that GQM structures may help to keep the questionnaire from getting bigger than it needs to be, by enabling the data managers to keep track of the data users present and past question (metric) needs.

The situation in the UI group was 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 eight 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 four laboratories of IBM's Software Solutions Division.

4.4.2 Impact on Data Usage

During the two years interval in which we observed the CUSTSAT MF, all DA/P modifications that originated from the existing process were done to improve running 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 were freely requested ad hoc, the GQM interviews did produce a few new improvement suggestions. This indicates that GQM interviews were an useful medium to get data users' feedback on data analyses and presentations (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 complementary to the other two. In the existing and GQM-based approaches, the user focused on improving the usage they are already making of the data. These approaches were driven by the immediate user needs. The AF-based method on the other hand pointed to new possible data usages. It was driven by the new insights that were gained from exploring the data.

4.4.3 Insights Gained from Data Analyses

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 during that year, the gap between IBM and competition had significant variations on two product satisfaction (PFSAT) attributes.

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. 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.

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

4.5 Cost Effectiveness

The total cost to improve and use the CUSTSAT MF inside the Toronto Lab was estimated to run between 500 and 800 person-hours (p-h) per year. This costs includes the current effort to modify and update the questionnaire as well as the effort spent with users groups doing data analyses and presentations. It does not include the day to day measurement framework operations and maintenance (e.g., data collection, database and data access maintenance, etc).

The total effort to apply the characterization (45 p-h), the AF data analyses (65 p-h), and GQM interviews (60 p-h) with three user groups was around 170 person-hours. These costs do include the data users and data managers efforts to learn the new methods, it does not include the time we spent to prepare and plan for this case study. In this scenario, we considered the new improvement approach 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.

In order to confirm this conclusion, we included in the subjective validation questionnaire a set questions aimed at evaluating the method's cost effectiveness. There were three questions for each phase of the new approach. There was one quantitative question using a five point scale and two open ended questions asking about the main benefits and drawbacks of the new approach methods. The quantitative question classified the method's cost effectiveness using a scale ranging from "(0) of no value" to "(4) of considerable value." The first question was about the MF characterization process. The data manager said that it was of "(2) of modest value." 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. This opinion is supported by our previous discussions. The characterization was a costly phase from the data manager point of view and Table 2 shows that it had little direct impact on the improvement objectives.

The GQM-based method was considered "(4) of considerable value" (highest score). 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 "(4) of considerable value" (highest score.) 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 better summaries of results. The second was that the obtained results usually need to be further explored statistically to prove their significance.

5 Concluding Remarks

We believe that an important problem in software engineering is to understand and improve existing measurement frameworks. Our work tackles this problem on two key fronts: 1) how to understand and structure ongoing measurement; and 2) how to better explore the data that the organization has already collected. We use the characterization step and the GQM-based method to tackle the first problem. We use the AF-based method 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). The case study tested the three parts of the new approach with respect to ten MF improvement objectives 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 prerequisite 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 our work is to show how the GQM Paradigm can be applied when the measurement framework is already operational.

The AF-based method describes how the AF technique can be applied in a measurement framework. The important result here is 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 Fig. 6). 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 (GRQs) and used to define AF analyses. The AF results can be fed back to measurement goals and used to revise existing GQM structures.

The new approach was designed to be nonintrusive to the MF management. 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.

The new approach is not by any means a complete or definitive approach to improve a measurement framework. It only contributes towards solving some of the problems that are associated with a MF. As seen in Section 4, it does have important capabilities that are complementary to the ones that already existed in the CUSTSAT MF. However, the methods we proposed have limitations that need to be addressed. The GQM-based method maps goals to metrics, the mapping from goals to data analysis and decision making models still needs to addressed. Our approach of applying the GQM-based method incrementally, one point of view at a time, has a limitation in that it does not detect *overall* extraneous metrics (i.e., metrics that are extraneous to the MF as a whole). 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 AF-based method mines data associations in nominal and ordinal data. The method should be further expanded to address other data mining techniques. The effectiveness of the AF-based method is dependent on the background knowledge and level of expertise of the domain experts that use it. It is our belief the likelihood of finding real, as opposed to spurious, knowledge during an AF data analysis session increases with the domain expert's background knowledge (expertise). Further research is needed to determine how sensitive the technique is from this factor. Unfortunately, our case study alone did not provide enough information to address the issue.

Lastly, the AF-based method is not a substitute for statistical data analysis techniques; it complements them. This method gives us the ability to find interesting facts that might otherwise remain hidden in the data. It is geared toward discovering information and supporting hypothesis formulation. One should use statistics to further analyze facts discovered through AF and, whenever possible, mathematically test the hypotheses raised during AF data analysis sessions.

It is important to highlight that bottom-up analyses may use data mining methodologies other than the AF-based method. Our choice of AF was determined by its simplicity, easy of use, and availability. AF's interestingness function is an association discovery technique. Other data mining techniques, be they based on association discovery, conceptual clustering, or sequence discovery, can in principle be used to develop methodologies to bottom-up analyze a measurement framework. We believe, however, that independent of the data mining technique chosen, a bottom-up analysis methodology should follow a process similar to the one described in Section 3.3 with well defined steps for data analysis planning, execution, and results interpretation.

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 immediate objectives are to better formalize the use of GQM to structure existing measurement frameworks, and combine it with different types of data mining approaches.

ACKNOWLEDGMENTS

This work was sponsored by the Center for Advanced Studies at the IBM Toronto Laboratory. We would like to acknowledge the support from the staff of the Center for Advanced Studies, Karen Bennet, in particular, and the Market Revenue and Planning Department, Jack Dawson, in particular, at the IBM Toronto Lab. We would also like to thank Inderpal Bhandari, formerly with IBM T.J. Watson and now with Virtual Gold Inc., for making this work possible with his data mining expertise. Last, but not least, we would like to thank the University of Maryland Experimental Software Engineering Group, Carolyn Seaman and Marv Zelkowitz, in particular, for their input at various stages of this work. Manoel Mendonça also recognizes the past support from CNPq (Brazil's Conselho Nacional de Desenvolvimento Científico e Tecnológico) for his research.

REFERENCES

[1] M.G. Mendonça, V.R. Basili, I.S. Bhandari, and J. Dawson, "An Approach to Improving Existing Measurement Frameworks," *IBM Systems J.*, vol. 37, no. 4, pp. 484–501, Nov. 1998.

[2] M.G. Mendonça, "An Approach to Improving Existing Measurement Frameworks in Software Development Organizations," PhD thesis, University of Maryland, College Park, MD, also available as CS-TR-3852 and UMIACS-TR-97-82, Dec. 1997.

[3] T. DeMarco, *Why Does Software Cost So Much?, Chapter 2: Mad About Measurement.* Dorset Housing Publishing, pp. 11–25, 1995.

[4] B.W. Boehm, J.R. Brown, and M. Lipow, "Quantitative Evaluation of Software Quality," *Second Int'l Conf. Software Eng.,* pp. 592–605, Oct. 1976.

[5] R.A. DeMillo and R.J. Lipton, "Software Project Forecasting," *Software Metrics,* A.J. Perlis, F.G. Sayard, and M. Shaw, eds., Cambridge Mass.: MIT Press, 1981.

[6] N.E. Fenton and A. Melton, "Deriving Structurally Based Software Measures," *J. Systems Software,* vol. 12, no. 3, pp. 177–187, 1990.

[7] A. Melton, D. Gustafson, J. Bieman, and A. Baker, "A Mathematical Perspective for Software Measures Research," *J. Software Eng.,* vol. 5, no. 5, pp. 246–254, 1990.

[8] H. Zuse, *Software Complexity: Measures and Methods.* deGruyter, 1990.

[9] A.J. Albrech and J.E. Gaffney, "Software Function, Source Lines of Code, and Development Effort Prediction: A Software Science Validation," *IEEE Trans. Software Eng.,* vol. 9, no. 6, pp. 639–647, Nov. 1983.

[10] L.C. Briand, V.R. Basili, and S. Morasca, "Property-Based Software Engineering Measurement," *IEEE Trans. Software Eng.,* vol. 22, no. 1, pp. 68–86, Jan. 1996.

[11] M. Shepperd, "Algebraic Models and Metric Validation," *Formal Aspects of Measurement,* I. Somerville and M. Paul, eds., Springer Verlag, 1992.

[12] J. Tian and M.V. Zelkowitz, "A Formal Program Complexity Model and Its Application," *J. System Software,* vol. 17, pp. 253–266, 1992.

[13] E.J. Weyuker, "Evaluating Software Complexity Measures," *IEEE Trans. Software Eng.,* vol. 14, no. 9, pp. 1,357–1,365, Sept. 1988.

[14] B.W. Boehm, "Software Engineering Economics," *IEEE Trans. Software Eng.,* vol. 10, no. 1, pp. 4–21, Jan. 1984.

[15] L.C. Briand, V.R. Basili, and W.M. Thomas, "A Pattern Recognition Approach for Software Engineering Data Analysis," *IEEE Trans. Software Eng.,* vol. 18, no. 11, pp. 931–942, Nov. 1992.

[16] C.F. Kemerer, "An Empirical Validation of Software Cost Estimation Models," *Comm. ACM,* vol. 30, no. 5, pp. 416–429, May 1987.

[17] C.E. Walston and C.P. Felix, "A Method of Programming Measurement and Estimation," *IBM Systems J.,* vol. 16, no. 1, pp. 54–73, Jan. 1977.

[18] V.R. Basili and R.W. Reiter, Jr., "Evaluating Automatable Measures of Software Development," *Proc. Workshop Quantitative Software Models,* Oct. 1979.

[19] V.R. Basili and D.H. Hutchens, "An Empirical Study of a Syntactic Complexity Family," *IEEE Trans. Software Eng.,* vol. 9, no. 6, pp. 664–672, Nov. 1983.

[20] V.R. Basili, R.W. Selby, and T.Y. Phillips, "Metric Analysis and Validation Across FORTRAN Projects," *IEEE Trans. Software Eng.,* vol. 9, no. 6, pp. 652–663, Nov. 1983.

[21] T.M. Khoshgoftaar, E.B. Allen, and D.L. Lanning, "An Information Theory-Based Approach to Quantify the Contribution of a Software Metric," *J. Systems and Software,* vol. 36, no. 2, pp. 103–113, Feb. 1997.

[22] N.F. Schneidewind, "Methodology for Validating Software Metrics," *IEEE Trans. Software Eng.,* vol. 18, no. 5, pp. 410–422, May 1992.

[23] B.A. Kitchenham, S.L. Pfleeger, and N.E. Fenton, "Towards a Framework for Software Measurement Validation," *IEEE Trans. Software Eng.,* vol. 21, no. 12, pp. 929–944, Dec. 1995.

[24] M.K. Daskalantonakis, "A Practical View of Software Measurement and Implementation Experiences within Motorola," *IEEE Trans. Software Eng.,* vol. 18, no. 11, pp. 998–1,010, Nov. 1992.

[25] R.B. Grady, "Successfully Applying Software Metrics," *IEEE Computer,* vol. 27, no. 9, pp. 18–25, Sept. 1994.

[26] S.L. Pfleeger, "Lessons Learned in Building a Corporate Metrics Program," *IEEE Software,* vol. 10, no. 3, pp. 67–74, May 1993.

[27] S. Rifkin and C. Cox, "Measurement in Practice," Technical report CMU/SEI-91-TR-16, SEI, 1991.

[28] E.F. Weller, "Using Metrics to Manage Software Projects," *IEEE Computer,* vol. 27, no. 9, pp. 27–33, Sept. 1994.

[29] T. Hall and N.E. Fenton, "Implementing Effective Software Metrics Programs," *IEEE Software,* vol. 14, no. 2, pp. 55–65, Mar. 1997.

[30] R.J. Offen and R. Jeffery, "Establishing Software Measurement Programs," *IEEE Software,* vol. 14, no. 2, pp. 45–53, Mar. 1997.

[31] W. Klösgen and J.M. Zytkow, "Knowledge Discovery in Databases Terminology," *Advances in Knowledge Discovery and Data Mining,* U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds., Cambridge, Mass.: AAAI Press/The MIT Press, 1996.

[32] N.E. Fenton, "Software Measurement: A Necessary Scientific Basis," *IEEE Trans. Software Eng.,* vol. 20, no. 3, pp. 199–206, Mar. 1994.

[33] V.R. Basili and H.D. Rombach, "The TAME Project: Towards Improvement-Oriented Software Environments," *IEEE Trans. Software Eng.,* vol. 14, no. 6, pp. 758–773, June 1988.

[34] V.R. Basili and D.M. Weiss, "A Methodology for Collecting Valid Software Engineering Data," *IEEE Trans. Software Eng.,* vol. 10, no. 6, pp. 728–738, Nov. 1984.

[35] V.R. Basili and S. Green, "Software Process Evolution at the SEL," *IEEE Software,* vol. 11, no. 4, pp. 58–66, July 1994.

[36] R.B. Grady, "Practical Software Metrics for Project Management and Process Improvement," *Hewlett-Packard Professional Books,* Chapter 3, 1992.

[37] J. Barnard and A. Price, "Managing Code Inspection Information," *IEEE Software,* vol. 11, no. 2, pp. 59–69, Mar. 1994.

[38] V.R. Basili, M.K. Daskalantonakis, and R.H. Yacobellis, "Technology Transfer at Motorola," *IEEE Software,* vol. 11, no. 2, pp. 70–76, Mar. 1994.

[39] K. El Eman, N. Moukheiber, and N.H. Madhavji, "An Empirical Evaluation of the GQM Method," *Proc. Center for Advanced Studies Conf. (CASCON 93),* pp. 265–289, Nov. 1993.

[40] 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 J.,* vol. 33, no. 1, Jan. 1994.

[41] 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 Trans. Software Eng.,* vol. 19, no. 12, pp. 1,157–1,170, Dec. 1993.

[42] 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 J.,* vol. 3, no. 3, pp. 167–189, 1994.

[43] I.S. Bhandari, M.G. Mendonça, and J. Dawson, "On the Use of Machine-Assisted Knowledge Discovery to Analyze and Reengineer Measurement Frameworks," *Proc. of Center for Advanced Studies Conf. (CASCON'95),* pp. 275–284, Nov. 1995.

[44] 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,* vol. 1, no. 1, pp. 121–125, Jan. 1997.

[45] E. Colet and I.S. Bhandari, "Statistical Issues in the Application of Data Mining to the NBA Using Attribute Focusing," *Proc. Am. Statistical Assoc. Section on Statistics in Sports, Joint Statistical Meetings,* pp. 1–6, Aug. 1997.

[46] Y.S. Lincoln and E.G. Guba, "Naturalistic Inquiry," *Sage,* 1985.

[47] L.C. Briand, V.R. Basili, and C. Hetmanski, "Developing Interpretable Models with Optimized Set Reduction for Identifying High-Risk Software Components," *IEEE Trans. Software Eng.,* vol. 19, no. 11, pp. 1,028–1,044, Nov. 1993.

[48] H. Potier, J. Albin, R. Ferreol, and A. Bilodeau, "Experiments with Computer Complexity and Reliability," *Proc. Sixth Int'l Conf. Software Eng.,* pp. 94–103, Sept. 1982.

[49] R. Selby and A.H. Porter, "Learning from Examples: Generation and Evaluation of Decision Trees for Software Resource Analysis," *IEEE Trans. Software Eng.,* vol. 14, no. 12, pp. 1,743–1,757, Dec. 1988.

[50] K. Srinivasan and D. Fisher, "Machine Learning Approaches to Estimating Software Development Effort," *IEEE Trans. Software Eng.,* vol. 21, no. 2, pp. 126–137, Feb. 1995.

[51] J. Tian and J. Palma, "Analyzing and Improving Reliability: A Tree-Based Approach," *IEEE Software,* vol. 15, no. 2, pp. 97–104, Mar. 1998.

[52] I.S. Bhandari, "Attribute Focusing: Machine-Assisted Knowledge Discovery Applied to Software Production Process Control," *Knowledge Acquisition J.,* vol. 6, no. 3, pp. 271–294, Sept. 1994.

[53] M.V. Zelkowitz and D. Wallace, "Experimental Models for Validating Technology," *IEEE Computer,* vol. 31, no. 5, pp. 23–31, May 1998.
[54] B.A. Kitchenham, L. P, and S.L. Pfleeger, "Case Studies for Method and Tool Evaluation," *IEEE Software,* vol. 12, no. 4, pp. 52–62, July 1995.
[55] B.A. Kitchenham, "Evaluating Software Methods and Tools. Parts 1 to 6," *Proc. AMC Special Interest Groups Software Eng.,* vol. 21, no. 1, Jan. 1996, vol. 22, no. 2, Mar. 1997.

Manoel G. Mendonça received his PhD in computer science from the University of Maryland at College Park in 1997. He also holds a MEng in computer engineering from the State University of Campinas (UNICAMP), and a BS in electrical engineering from the Federal University of Bahia (UFBa), both in Brazil. During the development of the work reported in this paper, he has been a visiting scientist at IBM Toronto Laboratory's Center for Advanced Studies. Until recently, he was a research associate at the University of Maryland and a scientist at the Fraunhofer Center for Experimental Software Engineering in Maryland. He is now a professor at Salvador University (UNIFACS) in Brazil. His main research interests are data mining, quantitative software engineering and management, software process improvement, and knowledge management in software organizations.

Victor R. Basili received his BSc in mathematics from Fordham College, Bronx, NY, MSc in mathematics from Syracuse University, NY, and his PhD in computer sciences from the University of Texas at Austin. He is a professor at the Institute for Advanced Studies and the Department of Computer Science at the University of Maryland. He is also the acting executive director of the Fraunhofer Center for Experimental Software Engineering in Maryland. His main interest are the development of quantitative approaches for software management, engineering, and quality assurance. He was one of the founders and principals of the Software Engineering Laboratory, a joint venture between NASA Goddard Space Flight Center, University of Maryland, and Computer Sciences Corporation. He has consulted with many agencies and organizations, and has authored more than 130 refereed papers. During his career, he has won several awards, including IEEE Computer Society's Outstanding Paper Award (1982), and Society of Golden Core (1997). He is the current editor-in-chief of the *Empirical Software Engineering Journal*, and has been a member of the IEEE Computer Society Board of Governors, editor-in-chief of the *IEEE Transactions on Software Engineering*, and program and general chairman for several conferences, including the Sixth and 15th International Conference on Software Engineering. He is an IEEE and an ACM Fellow.