A Goal Question Metric Based Approach for Efficient Measurement Framework Definition

Patrik Berander
Blekinge Institute of Technology
PO Box 520
SE-372 25 Ronneby
+46 (0)457 385840
patrik.berander@bth.se

Per Jönsson
Blekinge Institute of Technology
PO Box 520
SE-372 25 Ronneby
+46 (0)457 385815
per.jonsson@bth.se

ABSTRACT
In software engineering, measurements can be used to monitor, understand and improve software processes as well as products and resource utilization. Commonly, measurement frameworks are ambitious undertakings that require large data collection and analysis efforts. Unfortunately, such frameworks usually become too extensive, resulting in graveyards with a lot of data never being analyzed and used. One well-known way for defining measurements that tries to focus on the most important ones is the Goal Question Metric (GQM) approach. GQM focuses on eliciting goals and questions as drivers for finding metrics necessary to collect. This means that there is always a purpose with the defined measurements. GQM solves part of the problem mentioned, but it is still possible that the measurements defined become too many. In this paper, an extended GQM approach is presented that facilitates identification of and focuses on the most important measurements for an organization. The approach makes it possible to be more efficient when it comes to the application of measurement frameworks. Besides the presentation of the approach, two case studies are presented where the suitability of the approach is investigated in an industrial context. In the case studies, the suggested approach has been applied on two different processes within software development: change management and requirements engineering. From these case studies, a number of experiences are highlighted and discussed as issues to consider when applying the approach.

Categories and Subject Descriptors
D.2.8 [**Metrics**]

General Terms
Management, Measurement

Keywords
Goal Question Metric, GQM, Measurements, Requirements Management, Change Management

ISESE'06, September 21–22, 2006, Rio de Janeiro, Brazil.

1. INTRODUCTION
Software engineering organizations, as organizations in any other engineering discipline, need measurements to create a corporate memory which could be used to understand, control, improve, evaluate, and assess different aspects of the product lifecycle [4][14]. By using measurements, it is possible to get an overall picture of what is going on in an organization and determine the status and goodness of things as well as the organizational abilities regarding products, processes or resources [4]. By quantifying, it is possible to get hard numbers instead of relying on qualitative expert opinions which sometimes makes it hard to reach consensus about which decisions to make [4]. The importance of using measurements in software engineering is further supported by the fact that several quality models and process improvement methodologies (e.g. TickIT and CMMI) emphasize on measurements, and that there exist standards on how to conduct measurements (e.g. ISO/IEC 15939:2002 [6]).

Measurements involve, affect, and interest several stakeholders in an organization, which makes it essential to consider these stakeholders' opinions when defining measurements [4]. Measurements should also be aligned with the organization's direction, be linked to its goals, and reflect a balance of important perspectives and dimensions [11]. In addition, measurements should be used for beneficial rather than punitive purposes, and only those that actually will be used (the vital few) should be collected and analyzed [2][11][13].

Different approaches have been developed to guide the definition and implementation of measurement frameworks in order to avoid having too many measurements [5]. Going from goals to metrics has proven successful to assure relevant and used measurements [13]. The Goal Question Metric (GQM) approach, which was developed by Basili and Weiss and expanded by Rombach [12], is the most known goal-focused approach. GQM has become a de facto standard for the definition of measurement frameworks [12], despite some reported weaknesses [5][10]. One of the reasons for its success is that it is adaptable to many different organizations and environments, as confirmed by the large number of companies that have used it (e.g. Philips, Siemens, NASA) [14].

Another reason for the success of GQM is that it aligns with organizational directions and goals. This is done by focusing on measurement goals in a top-down fashion instead of focusing on a metric-driven bottom-up approach [14]. Also, the measurements are elicited and defined by the experts in the organization, i.e. the employees [14]. The application of GQM results in a measurement tree containing the following three levels [14]:

- The **conceptual level** (Goal) contains measurement goals regarding products, processes, or resources, specifying the objectives of the measures.
- The **operational level** (Question) contains a set of questions used to characterize how the goal(s) should be attained.
- The **quantitative level** (Metric) contains objective or subjective quantitative metrics that are collected in order to answer the question(s).

In the tree structure, the goals, questions, and metrics are related, sometimes with many-to-many relations between each other (e.g., one question could be related to several goals). The definition is conducted top-down (i.e., goals generate questions, which in turn generate metrics) and the interpretation/analysis is conducted bottom-up. Using GQM means that the measurements are collected for a purpose and hence also reflect interesting and useful measurements for an organization, instead of just easy ones [4]. Despite this, and despite that GQM is said to reduce the number of measurements, several authors (e.g., [10][13][15]) report that there is a risk that more measurements are identified than are possible to collect and analyze. One factor of defining successful measurement frameworks is to start small with the most important measurements and grow slowly as the organization matures, especially if measurements are being tried for the first time [6]. Otherwise, there is a risk of getting overwhelmed with data. If it is not known which measurements that are important, there is a risk that no measurements are collected and analyzed because it is not clear which ones to focus on (or that the "wrong" measurements are collected which is a waste of effort) [4][10]. After all, it is better to use a few useful measurements than none at all.

Since the original version of GQM does not explicitly support limitation of measurements to a reasonable amount, there is a need for extending GQM to take this issue into account. This paper suggests an extended GQM approach that inherits all the positive aspects from GQM while using prioritization as a tool for facilitating the limitation of measurements. Besides the limitation support, it also allows for balancing different dimensions through categorization. The approach also increases the involvement of organizational personnel and hence should be able to boost commitment. These extensions to GQM makes the approach more aligned to ISO/IEC 15939:2002, where it is emphasized that measurements shall be prioritized by different stakeholders and only a subset of identified measurements shall be selected to make the measurement framework efficient (see [6]).

The approach presented in this paper has been developed with industrial needs in mind. It has been tested in industry in two different case studies and has improved based on industrial feedback and experiences. In the case studies, the aim was to find and define as few measurements as possible while still obtaining organizational goals (also referred to as a lean measurement framework [8]). In its original form, GQM is divided into four separate phases (planning, definition, collection, and interpretation) where the elicitation and definition of measurements are performed in the definition phase [12]. The approach suggested in this paper focuses only on the activities in this phase and it is possible to conduct the remaining three phases as usual. Further, it should be noted that the suggested approach is a conceptual process of how to define goals, questions, and metrics in an efficient way. This means that it is possible to use variants of the GQM approach, seen in for example GQ(I)M [9] and [3]. Further, for the part of the process that this approach covers (definition), it aligns fully with the ISO/IEC 15939:2002, in which it is stated that measurements shall be identified, prioritized, selected, and documented.

This paper is structured as follows. In section 2, the extended GQM approach is presented. Section 3 presents two case studies that have been conducted and that have served as input to the suggested approach. Section 4 reports the overall experiences while Section 5 concludes the work performed and the findings.

2. EXTENDED GQM APPROACH

Having too many measurements to collect and analyze might imply that no measurements are used at all. To cope with this, a structured approach to find usable and useful measurements for different processes (processes are in focus in this paper but the approach is not specific for this use) is presented. This approach is based on the Goal Question Metric (GQM) approach (see Section 1) but is extended to address drawbacks with GQM. Furthermore, the approach has been incrementally refined and has matured based on the experience gained from two consecutive case studies (see Section 3) in an industrial setting. The resulting approach based on this experience is visualized in Figure 1.

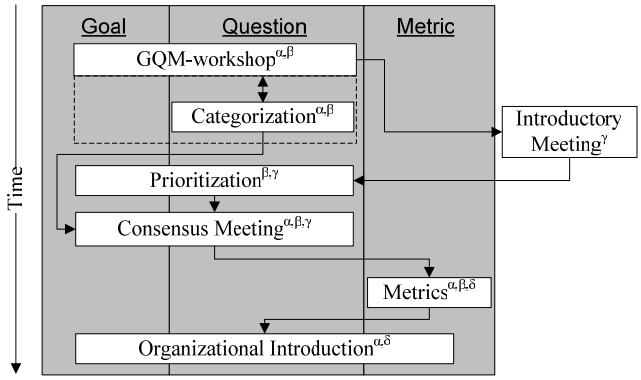

Figure 1. Work process of the extended GQM approach.

Figure 1 is divided into three areas (i.e., Goal, Question, and Metric) where the steps in the approach (e.g. categorization) are placed. The areas overlapped by each of the steps represent which parts of GQM that are covered in the particular step (e.g., prioritization is performed at both goal and question levels). In relation to each step, the following symbols are used to indicate which people are involved in the particular step:

- α: GQM team (moderator)
- β: Organizational representatives in the GQM workshop
- γ: Organizational representatives excluding the ones involved in the GQM workshop
- δ: Receiving role within the organization

The steps are presented in the subsequent sections.

2.1 GQM-workshop

The first step is to identify goals and questions interesting for the organization at hand, in one or several workshops. The persons representing the organization should have suitable roles, knowledge, and insight in the work process being studied. They should also

have recent "hands-on" experience from project work, meaning that operative personnel are favored over functional (line) personnel. It is recommended to not include too many persons since it can make the workshop inefficient. However, the main perspectives in the work process must be represented.

Besides the participants from the organization, a neutral GQM team, consisting of one or several persons, is essential. The members of the GQM team act as moderators and should have good knowledge of the GQM approach. It is also preferred that they have some knowledge about the work process as well as the organizational culture and language.

When conducting the workshop, the moderator starts by introducing GQM to the participants (assuming the participants have no prior knowledge of it). It is good to present the purpose of measurements, how GQM works and the ingredients of it, and how the work is to be performed. It is also advisable to give examples of goals, questions, and metrics to quickly get the participants to understand their task. However, the examples should not be from the process in focus since that may bias them.

The goals should be identified first, followed by the questions. However, GQM is a highly iterative process where goals often are identified when working with questions. The moderator should also be aware that participants may try to formulate improvement goals rather than measurement goals (measurement goals specify the objectives of a measurement framework [14]).

The moderator should have an active role to prevent some participants from being overly influential. Some will talk more than others and the moderator should assure that all participants suggest goals and questions. It should also be assured that participants are not afraid of having ideas and suggestions. No goals or questions should be rejected because they are "stupid" since important aspects might then be lost. Unsuitable or unimportant goals and questions are instead purged later if others think that they are not suitable (see Section 2.5).

To get an overview of the process in focus, and to make sure that different parts of this process are covered, it may be a good idea to visualize the process by relating goals and questions to the different process steps. This makes it possible to see if any parts of the process are missed out and should be complemented.

The time spent on the workshop is dependent on the topic, the participants, the organization, etc. The moderator should be able to determine when the participants do not contribute with any more "real" goals or questions, for example when they repeat themselves or discussions become off-topic.

2.2 Categorization

Measurements often represent different dimensions and are collected for different purposes. When later (see Section 2.5) choosing among the questions identified, it is easy to only focus on issues that are most important at the moment. Categorizing the questions according to their characteristics makes it possible to consciously include questions for several dimensions and hence minimize the risk for ending up with questions only covering a few dimensions. Categorizing also provides guidance and context, making it possible to be proactive and make sure that some dimensions are not missed when eliciting goals and questions (in which case the categories of course must be available in advance).

The categorization activity should be performed in conjunction with the GQM-workshop as shown by the dashed line in Figure 1. While it is a separate step of the approach, there is a feedback connection to the workshop activity.

The categories used in the categorization can, roughly speaking, come from two sources. Either they have been defined before the GQM-workshop, or they are defined during the workshop based on the elicited questions. In both cases, the output of the categorization can be used for determining if more questions need to be elicited. The main difference is that predefined categories allow us to see whether all dimensions are covered when eliciting questions, whereas categories defined based on the questions rather indicate the balance of questions among categories. It should be noted that the categorization is important also during the consensus meeting when choosing among goals and questions to include (see Section 2.5).

2.3 Introductory Meeting

When performing the prioritization (see Section 2.4), it is a good idea to increase the number of participants to get a broader organizational anchoring than if only involving the participants of the GQM workshop. It is then possible to involve different roles and perspectives to account for different opinions about goals and questions. Before the prioritization, these participants should get an introduction to the task and the possibility to ask questions. The introduction should include a short presentation about how GQM works, an overview of the elicited goals and questions, information about how the prioritization should be conducted, as well as information about tools used for prioritization.

2.4 Prioritization

Goals and questions identified are seldom equally relevant, urgent, or important. It is also likely that more are found than are possible to realize within a reasonable amount of time and cost. One way of reducing the number of goals and questions is to prioritize them. In this approach, Hierarchical Cumulative Voting (HCV) is chosen for prioritization because it works on a ratio scale (all arithmetic operations can be applied), it can cope with hierarchically classified objects, and it makes it possible to care for different stakeholders' views [1]. Multiple stakeholder support is important as it is recognized as important to involve different stakeholders in the prioritization process [6]. Supporting different levels is important because it gives the possibility to prioritize both on goal and question levels, while also reducing the number of objects to prioritize at a time. The use of HCV in a GQM environment is explained below and illustrated in Figure 2.

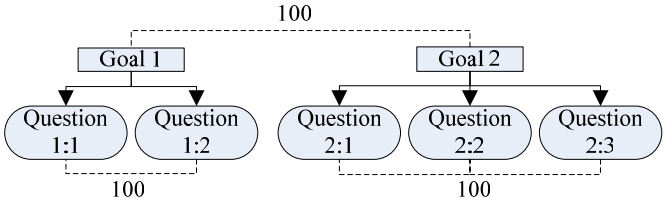

Figure 2. Example of prioritization.

HCV is based on that the prioritizing person gets 100 imaginary units (dollars, points, etc.) to distribute between the different objects to prioritize. When having multiple levels, as when prioritizing goals and questions, every stakeholder is first given 100 points to

distribute between the different goals (Goal 1 and 2 in the example) and then 100 points to distribute between the questions within each goal.

When all goals and questions are prioritized, the final priority for each question is calculated by multiplying the relative weight of each question (within its goal) by the relative weight of the goal. When doing this calculation, a compensation factor should be utilized due to the nature of the hierarchy (see Section 4.4.4) to compensate for different number of questions within the goals (see [1] for more details about this calculation). By prioritizing on both goal and question level, it is possible to later focus on goals and questions separately or in combination (see Section 2.5).

After each individual prioritizing person's result has been collected, an overall result is calculated by summing together the result of all prioritizing persons. Here, it is also possible to assign weights of importance on the participants and hence give different participants a different amount of influence on the end result. Consult [1] for details about how to use HCV.

2.5 Consensus Meeting

Based on the priorities, a subset of goals and questions shall be selected to be implemented in the measurement framework. This is done in the consensus meeting, which is useful for three primary purposes:

- Determining which goals and questions to include
- Discussing the result
- Sharing experiences (knowledge transfer)

By focusing on the first purpose, the other two are fulfilled as a bonus. At the meeting, it should be discussed if the priorities of goals and questions are reasonable given the whole picture. Next, it should be determined which goals and questions to include. Some example strategies for determination are presented below:

- Pick a fixed number of goals and/or questions to include (e.g., the 5 highest prioritized questions).
- Agree on a relative weight that must be reached (e.g. questions with higher priority than 10 percent, or the highest prioritized questions until 50 percent of the importance is reached).
- Include the highest prioritized questions until all categories are covered, or pick the highest prioritized questions from each category.
- Include all questions related to one or more goals.

Any of the above strategies or combinations thereof can be used. In addition to the priorities, it is important to consider additional criteria when choosing which measures to select. Examples of such criteria are: dependencies, feasibility, availability, cost, and ease of collection (see [6] for further examples). The participants of the consensus meeting (same persons as in the prioritization) should jointly determine what is important to consider and what strategy to use. It is also a good idea to prepare this meeting by distributing the results in advance to the participants. The results should then be presented in different ways (ordered lists, Pareto charts etc.) to account for individual preferences.

2.6 Metrics

The next step is to identify metrics for the chosen goals and questions. This should be done by the persons participating in the GQM-workshop, members of the GQM team, the person(s) that will be responsible for the measurement framework in the future, and possibly also experts on specific metrics. The GQM team hosts this part of the approach and the primary focus is to connect metrics to the questions. By involving the originators (workshop participants), the receiver(s) (responsible for the measurement framework), and experts regarding certain areas concerned, knowledge is shared between the parties.

2.7 Organizational Introduction

After the previous steps, it is important to document and communicate the results to the stakeholders [6]. The information to document concerns organization-specific issues such as who will collect the metrics, who will analyze the metrics, when and how the metrics should be collected and analyzed, and expected results of the measurement. It is also important to determine what needs to be done before it is possible to collect and analyze the metrics (e.g. changes in processes or tools). This work is preferably performed by members from the GQM team together with the future responsible of the measurement framework. Note also that it may be of interest to report the goals and questions not chosen for implementation, since they may be introduced later [6].

3. CASE STUDIES

The final approach presented in Section 2 has been developed and refined through two industrial case studies, which are presented in this section. Thus, both case studies presented below differ somewhat from the approach presented in Section 2. Section 4.1 highlights differences and explains how the approach evolved based on the lessons learned.

3.1 Case Descriptions

Both case studies were conducted at one product development unit at Ericsson AB, an ISO 9001:2000 certified company, in Sweden. Ericsson is one of the market leaders within the telecommunications domain and sells systems to a market with basically all mobile operators as potential customers. The development unit has about 400 employees working in different software development projects (primarily real-time telecommunication systems). The projects typically include 60-120 persons for 12-18 months. The product management organization elicits and specifies high-level requirements that are broken down and refined within projects for further design, implementation and testing. Since the projects last for almost two years, and because the telecommunication business is moving very fast, the projects are subject to changes. Changes to the projects are handled by a change management process similar to the generic process described in literature related to change management of requirements (e.g., as described in [7]).

The approach has been evaluated with two different processes, one in each case study. These processes were chosen together with management based on an evaluation of important improvement areas. The processes in focus were the change management process and the requirements engineering process, respectively. In both case studies, the GQM team consisted of two persons (the authors of this paper) who acted as moderators and cared for the documentation of the results. Both moderators had good knowledge about the organization, work language, etc. due to previous collaboration. By

using two moderators, it was possible to share the responsibility (even though one was considered as the main moderator) and one moderator could catch up on issues missed by the other moderator.

The upcoming two sections present the details and results of the two case studies. The approaches presented in these sections differ from the approach presented in Section 2. The reason for the deviations is that the approach presented in Section 2 has matured based on the experiences gained from these two case studies.

3.2 Case 1 – Change Management

The first case study was conducted with the change management process in mind. The approach used to elicit and limit the measurements is presented in Figure 3 and the details of each step are presented in the subsequent sections.

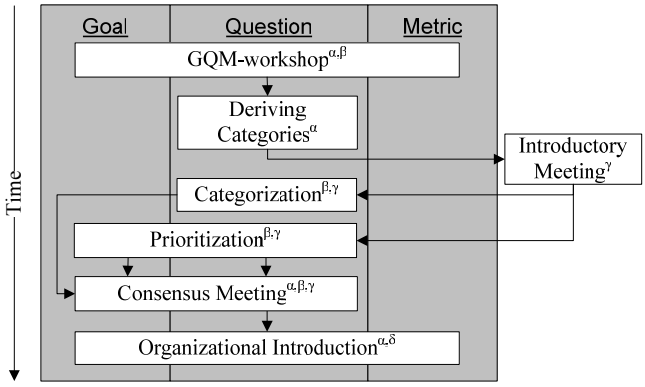

Figure 3. Work process in case 1.

3.2.1 GQM-workshop

Three persons involved in change management at Ericsson were invited to identify goals, questions, and metrics. They represented the roles chairman of change control board, project manager, and technical coordinator. One of the moderators was responsible for continuously translating the discussions into goals, questions, and metrics on a projector-aided computer (using Microsoft Visio) which made it possible for the participants to see the result all the time. The moderators' most important task was to make sure that everyone's thoughts were captured; not only those of the ones who talked most and loudest. Four workshops were held, summing up to nine hours per person. Between each workshop, the researchers prepared the next one by refining and organizing goals, questions, and metrics, as well as identifying possible weak spots. Two of the nine hours were devoted to validation, i.e. checking that the translation from Swedish to English (made by the GQM team), as well as the various refinements, were correct and that nothing was missing. Further, half an hour of the first workshop was devoted to introduction of people and GQM by the main moderator. The result from the workshops was that 7 goals, 24 questions, and around 40 metrics were identified.

3.2.2 Deriving Categories

With goals, questions, and metrics identified, the questions were evaluated from a categorization perspective by the GQM team. Categories were derived by looking for common denominators among the questions. Six categories were identified that together covered the dimensions and purposes of the questions involved:

- Process adherence (how well the process is followed)
- Environmental impact (impact on the project/product/organization by uncontrollable internal and external factors)
- Internal impact (how exposed different parts of a project/product/organization are to changes)
- Effectiveness (how good/effective the process is)
- Characteristics (value-neutral information about some aspect of a project/product/organization)
- Snapshot (provides a picture of a project/product/organization at the moment of measure)

The explanations above are summaries of the full category definitions. It should be noted that they are not a final list of categories but were derived in this particular context and based on the view of the members of the GQM team.

3.2.3 Introductory Meeting

The next step was to let a larger sample from the organization categorize the questions and prioritize the goals and questions. The GQM workshop participants identified additional persons (a representative sample of roles) to invite to an introductory meeting and for the categorization and prioritization. At this meeting, the task was presented (background, goals and questions, how to prioritize and categorize) by the main moderator. The attendants did not have any objections or problems of understanding, but they identified some additional persons that were considered as important to include in the prioritization and categorization. After this meeting 19 potential respondents were invited to do the prioritization and categorization.

3.2.4 Categorization

The categorization was made by letting the respondents rank the suitability of the categories in relation to each question. This way of establishing categories was tried out in two pilot categorizations, one in the workshop group and one in an academic environment with seven researchers. Based on the experiences from the pilot studies, it was decided to send out the categorization as a questionnaire electronically (using Microsoft Excel). This reduced the risk of errors (by setting limitations for valid inputs) and made it highly efficient. An added benefit was that the analysis of the results was made easy.

Out of the 19 invited, 15 persons performed the categorization as intended. In some questions the participants agreed very much while they agreed less in others. "Characteristics" was the category assigned to most questions (8) while "Environmental impact" was the least chosen one (1). The stakeholders also had the opportunity to state that neither of the predefined categories was applicable, but this option was not used.

3.2.5 Prioritization

Microsoft Excel was used for the prioritization (for the same reasons as in Section 3.2.4), which made it possible to guide the participants in the process. Each tab in the workbook contained the goals or questions to prioritize, and the 16 participants that did the prioritization as intended continuously got information about how much points they had left to distribute. If exceeding the allowed number of points, they got a message telling them to decrease the number of points.

It was not possible to see any patterns of preferred categories in the prioritization (i.e., no category was more important than the others) or any sign that specific roles preferred particular questions. Since there were 24 questions to prioritize, the average priority of the questions was slightly more than 4 percent (100/24). Out of the 24 questions, 9 got a priority higher than this value. The question that got the highest priority got 11 percent while the question with the lowest priority got 1 percent.

3.2.6 Consensus Meeting

Twelve persons participated in a one-hour consensus meeting. An e-mail was sent out prior to the meeting with the results of the prioritization and categorization. At the meeting, the result was presented by the GQM team and the participants had a chance to ask questions about the result. The next step was to discuss if there were any questions that were considered as having incorrect priorities. One of the participants argued that some of the questions had too low priorities. These were questions that aimed at finding reasons for change requests, and hence facilitate improvement of for example requirements elicitation. The other questions were more focused on the change management process as such (e.g., number of wrong decisions and effectiveness of decisions). It was decided to divide the questions into two groups:

- Change root cause questions
- Change process questions

Both these kind of questions add value to the organization, even though they have different stakeholders (especially when it comes to the analysis and the actions taken from the results). Most of the participants had not thought about change root cause questions as important previously, but agreed on their importance. The participants considered them as equally important as questions regarding the change process. The result of this was that some of the questions considered as less important in the priority list, were considered as essential after the discussions. This shows the added value of having a consensus meeting where the result is discussed. The participants agreed to find the most important questions for both types of questions.

From the change process perspective, the highest prioritized questions were chosen until the participants thought the needs were fulfilled. The four highest prioritized change process questions were chosen (belonging to the categories "Effectiveness" and "Process adherence") with a total of 35 percent of the importance. The change root cause questions were more difficult since they had not been prioritized correctly before. After some discussions, it was agreed that all of these questions should be included (belonging to the categories "Internal impact", "Characteristics", and "Environmental impact") with a total of 28 percent of the importance. Hence, the chosen questions together accounted for 63 percent of the importance, while only choosing 46 percent of the number of questions for implementation. All categories except for "Snapshot" were included in the final result. This shows that the thought behind the categorization was followed and that several perspectives were accounted for in the questions.

The participants thought it was really important to continue working with the resulting measurements. They were really eager to see the result of the work and were determined that the framework should be introduced in the next upcoming project.

3.2.7 Organizational Introduction

The final step was to transfer the results back to the organization. This was done by the main moderator together with the person responsible for the change management process (and additional discussions with the GQM workshop participants). The following information was defined and documented:

- For goals: Slogan, purpose, priority
- For questions: Slogan, priority, stakeholders, type and category
- For metrics: Slogan, when it should be collected, role responsible for collecting the metric, attributes needed to collect the metric, tool requirements put on the process specific tool, current tool compliance, example of results (such examples were basically the indicators in the GQ(I)M approach as discussed in [9])

In addition, a document was produced where it was defined which attributes the change requests must have in order to enable collection of the metrics (and hence answer questions and fulfill goals). The documents produced were then used as a driver for process improvement, as a foundation for putting requirements on the tool vendor, as well as guiding the work with the measurement framework in the future. This way of defining and transferring the results was useful and appreciated within the organization.

3.3 Case 2 – Requirements Management

The second case study was done in relation to the requirements engineering process at Ericsson. Based on the experiences gained from the first case study, parts of the approach were modified to make the work more efficient. The work process used is presented in Figure 4, and each step is reported in the subsequent sections.

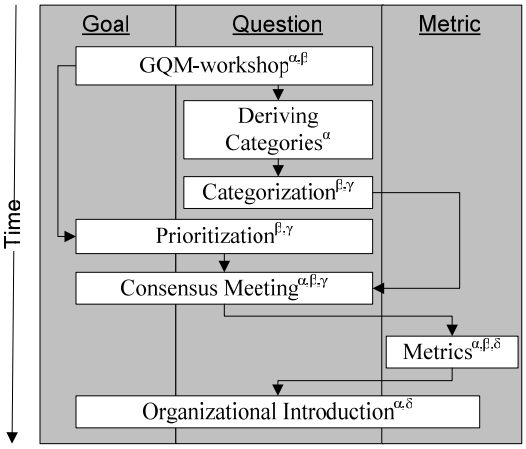

Figure 4. Work process in case 2.

3.3.1 GQM-workshop

Four persons involved in requirements management at Ericsson were invited to the workshop. The roles represented were system designer (responsible for the design of the system), operative product manager (customers' spokesperson in a project), tester, and technical coordinator. The workshop was conducted similarly as described in Section 3.2.1, except that metrics were not identified. The reason for this was to avoid spending effort on identifying metrics for questions that should be excluded later on.

The identification was conducted in three workshops summing up to five hours per person where half an hour of the first workshop was devoted to introduction of people and describing GQM and one hour was devoted to validation (plus a final validation per e-mail). The result of the workshops was six goals and 25 questions (where six of the questions were shared between two goals).

3.3.2 Deriving Categories

The work with deriving categories was conducted by the GQM team. However, when comparing the questions elicited in the workshop with the categories defined in the previous case study (see Section 3.2.2), it became clear that those categories could be reused (with some minor modifications) in the requirements process as well. In other words, all categories contained questions and each question could be placed in a category.

3.3.3 Categorization

The categorization was conducted in the same way as presented in Section 3.2.4, except no pilot categorization was done. 22 persons (11 roles) were asked to perform the categorization and 18 performed the categorization according to the instructions. All categories were chosen at least once and no questions were considered as not being represented by the categories available. The extent of agreement regarding categories varied between questions. "Internal impact" was the category assigned to most questions (6), "Snapshot" was the least chosen one (2), and no questions were seen as having no suitable category. The questions that were shared between two different goals were categorized separately in the context of each goal (where two questions ended up in different categories depending on the goal).

3.3.4 Prioritization

The prioritization was conducted in the same way as presented in Section 3.2.5. Of the 22 persons invited, 19 did the prioritization as instructed. It was not possible to find any patterns of preferred categories (i.e., no category was more important than others) or role differences in the prioritization. The average priority of the questions was 4 percent (100/25) and nine got a priority higher than this value. The highest prioritized question got 11 percent while the lowest prioritized got 2 percent. One goal clearly dominated the result; as it was prioritized as representing 32 percent of the importance.

The questions that were shared between two different goals were prioritized under both goals separately, mainly because it was argued that a question could vary in importance depending on context (i.e., goal).

3.3.5 Consensus Meeting

All persons answering the prioritization and categorization were invited to the consensus meeting and eight of these came. Since the meeting was held during flu period, the number of participants was limited. The participants were given the result both in tables and Pareto charts prior to the meeting. At the question level, the participants were given the final priority list presented both with the shared questions separated and added together. Although the participants agreed that it made sense to look at both lists, they finally decided to look primarily at the list where the priorities were separated for shared questions.

At the one-hour consensus meeting, nobody thought that any goals or questions should have different priority than they had been given in the prioritization. Instead, the discussions focused on which goals and questions to choose. Several different variants of choosing the goals and questions were discussed. It was quickly decided that the highest prioritized goal (and its questions), representing 32 percent of the importance, should be chosen (the questions were ranked as 1, 2 and 4 at the question level). As the questions related to the highest prioritized goal were considered as fairly expensive, it was finally decided that only these questions, together with the question ranked as number three should be chosen (they regarded this third ranked question as cheap to answer while still providing much value). The four out of 25 questions chosen (16%) represented three categories ("Characteristics", "Effectiveness", and "Environmental impact") and accounted for more than 31 percent of the importance.

3.3.6 Metrics

Even though the participants at the consensus meeting were really interested in being a part of the definition of the metrics, it would be hard to have them all discussing and defining metrics. Instead, the participants of the GQM-workshop together with the GQM team and the person later responsible for the measurement framework were selected to meet in a one hour session to define the metrics. At this meeting, only one of the questions was handled. The main reason was that a major process change was introduced at the time, and many discussions focused on how things would work with this process. Even though these discussions were very fruitful from an organizational point of view, only metrics for the particular question were defined. Due to this, another meeting was scheduled for the definition of the remaining questions. Besides the persons involved in the first session, it was decided to invite experts with roles needed for defining the remaining metrics in a good way. At the second session, one of the remaining questions was fairly easy to define suitable metrics for (not at least due to the experts invited) while the last two were harder. The reason for having problems with these was that some process improvements were needed before it was possible to make use of these metrics. Since these two questions were the highest prioritized, it was decided that these process improvements were crucial to do. This second session took almost two hours and much of the discussions were focused around how to perform the process improvements in order to be most cost efficient.

3.3.7 Organizational Introduction

The organizational introduction was performed by the GQM team and by a person responsible for the requirements process. This was done similarly as presented in Section 3.2.7. In addition to preparing a report as a driver for introduction, a presentation about the results was prepared and presented to a management group within the organization. Here, special attention was given to the questions that were not possible to find metrics for until some process improvements were in place. The management group was very interested in these results and immediately started a work group to get these process improvements in place. One of the main drivers for triggering this immediate start of the process improvement was the fact that the management group saw the high priority of the goals and questions, and realized that it was a highly prioritized area. This way of introducing the results was again very much appreciated.

4. EXPERIENCES

In this section, experiences from the two described case studies are discussed. The experiences are divided into subsections according to which phase of the approach they belong to. First, however, the

evolution of the final approach through the two case studies is described.

4.1 Evolution of the Approach

In case study 1 (change management), metrics were elicited as a part of the initial GQM-workshops as is normally done in GQM. However, as the purpose of the suggested approach is to limit the number of measurements to the most important ones, many of the elicited and discussed metrics were excluded later in the process. Hence, the work done with excluded metrics could be seen as wasted effort. Therefore, the definition of metrics during the GQM-workshop was seen as unwise, and it was decided to make it a separate, later, activity. This way of handling the metrics is also proposed in ISO/IEC 15939:2002, where it is stated that information (goals and questions) may be identified but not fully defined (metrics) or implemented until required [6].

In both case studies, the categories were derived by the GQM team alone instead of together with the workshop participants. It was later argued that this activity should be done in the workshop in order to get the participants involved and get their input. Also, categorization was in both case studies performed by the persons who did the prioritization. Two problems exist with this. First, it is not possible to get feedback during the workshop regarding whether or not all categories are covered. Second, the persons doing the prioritization considered the categorization activity to be difficult and argued that it would be easier to do it as part of the workshop. Thus, it was decided to have the categorization earlier, as a part of the GQM workshop.

In case study 2 (requirements management), no introductory meeting was held before prioritization and categorization because it was not deemed necessary. However, the meeting was included in the final approach again due to requests from the case study participants.

4.2 Workshop

4.2.1 Moderation

As described previously, there were only a few participants in the workshops in addition to the GQM team; three in the first case study and four in the second. This resulted in a very open workshop climate, where the participants could freely discuss ideas and opinions (see [13]). The moderation of the workshops was kept at a low level to avoid confining the discussions and limiting ideas. The effect was that sometimes some workshop participants became more vocal than others, and thus overly influential. Moreover, off-topic discussions that eventually lead to no results were occasionally witnessed. In summary, it is seen as important that a moderator is able to make correct trade-offs regarding when to let discussions flow, and when to interrupt.

4.2.2 Mindset

The mindset of the workshop participants seems to play a significant role when it comes to efficiency in creating relevant goals and questions. As stated previously, the goals resulting from the GQM application should be measurement goals, i.e. have clear measurement focus. Despite this fact, the workshop participants sometimes formulated abstract improvement goals, or started discussing general improvement issues instead of focusing on measurement goals. However, it is sometimes possible to derive measurement goals from such improvement goals, at the same time as measurement goals can support the organization to achieve improvement goals [14]. In addition to the case studies presented in this paper, GQM has been tested with students as a part of an industrial development project and these students had no problem with formulating concrete measurement goals. This is probably because of different mindsets; the participants in the case studies were biased by their relationship to the organization in which they worked, while the students were more "free-thinking", and could quicker adapt to the method. One possible way to facilitate the elicitation of goals and questions may be to provide a "reference GQM tree" that could help the participants with an appropriate mindset. However, caution should be exercised to prevent the reference tree from affecting the result.

4.2.3 Continuation

In both case studies, multiple workshop meetings one to two weeks apart were held. An effect was that each meeting started with revisiting the material produced in the last meetings, in order to refresh the memory. This took some time but, more importantly, it also lead to that the material was updated (e.g. something was added or rephrased) from time to time. An obvious advantage of this is that the material produced can gradually improve, but there is also a risk that previous thoughts and ideas are misinterpreted in the update process which means that the original meaning is lost. This risk could of course be minimized by formulating carefully directly. An interesting observation is that the workshop participants often saw formulations as being final, whereas the GQM team cared for finding the essence of the discussions and later clarify, complete or improve. In summary, using multiple workshops is a good way of letting ideas mature, but they should not be too many resulting in tampering the initial material too much.

4.3 Categorization

4.3.1 Who and When

In the case studies, the categorization of questions was done in parallel with the prioritization, by the same persons. From this activity, data were collected about how understandable the categories were, how easy it was to categorize and to what extent the categories were mutually exclusive. It was found that the prevailing opinion among those who did the categorization was that it was rather difficult to assign categories and that many questions ended up in multiple categories. While the categorization seems to be a good way of knowing if the set of questions is complete in some aspect, it is clear that it is difficult to get consistent categorization results when it is not the same people who define the categories that perform the categorization. Further, the categorization was also perceived as harder in case study 2 without an introductory meeting which indicates that the problem partly was related to knowledge about the questions and categories. This motivates making the categorization a part of the workshop (as can be seen in Section 2.2), in which case both the definition and the assignment of categories would be performed by the same persons. An added benefit would be that it is possible to get quicker feedback on the completeness of the questions.

4.3.2 Relevancy

As previously stated, categorization is a way of determining the completeness of the question set. This brings up the issue of defining a useful set of categories. Usefulness can in this context be divided into completeness and goodness, with the former being the extent to which the category set captures all relevant types of questions, and the latter being the extent to which a category is

meaningful and mutually exclusive. It is likely that there are two types of categories, generic and company specific (or domain specific), both of which should be addressed. In the two case studies, the categories used were only of the generic kind, and the lack of company-specific categories may be explained by the fact that the ones who defined the categories come from academia. That might also explain why the persons who categorized the questions sometimes had problems relating to the categories, and further indicates that both category derivation and categorization should be done as a part of the GQM workshop.

4.4 Prioritization

4.4.1 Focus

When discussing the prioritization with people from the organization, it became clear that there is a difference between considering measurements to be interesting and considering them to be important. For example, one person stated that while he could collect measurements, it did not mean that he could (or had time to) use them to improve later projects. The GQM approach is designed to produce metrics that exist for a purpose, but the organization still needs to have a process in place for making use of the data collected. Thus, it is imperative that those who prioritize questions do it from the point of view of importance rather than mere interest. In other words, the questions that are ranked as most important should be the ones for which measurements are most likely to be made use of. The only rational solution is to specify really well from what aspect (i.e., importance) the questions should be prioritized.

4.4.2 Importance vs. Cost

During the consensus meeting for the second case study, the issue of looking at both cost and importance for the measurements was raised. Looking at a cost perspective for the measurements would simplify the selection of relevant measurements, as those with very high cost (in compared to their importance) could be easily dismissed. Determining the cost of a measurement can be difficult, but one possible way is to consider if it is already handled (possibly in a different form), if it only requires a small addition to a process, or if a new process needs to be established.

4.4.3 Dependencies

Due to the nature of the GQM approach, there are of course vertical dependencies between questions and goals. However, horizontal dependencies may also exist in the GQM tree. One type of horizontal dependency between questions occurs when the questions are descendants of the same goal. Questions related to a particular goal contribute to it and can hence be seen as having value dependencies between each other. In the consensus meetings (especially in case study 2), this dependency type became visible as the participants reasoned that if they took one high-ranked question, they might as well take another one contributing to the same goal, even if it was much lower ranked. Another dependency type occurs when multiple goals share the same questions, or when multiple questions share the same metrics. Intuitively, the situation is easy to deal with, as the extra goals or questions can be seen as an additional bonus you get when collecting measurement data and answering the questions. In practice, however, the situation becomes more complex. For example, should a question that is shared among several goals be prioritized once, or one time for each goal? In the latter case, should the separate question priorities be used, or should they be combined to one single priority? This type of dependency was present in the second case study, where two goals shared the same set of questions. Each of these questions was prioritized twice, and the results were discussed both with and without summing the priorities of the questions together. While the results differed much, not enough data have been collected to favor either way of dealing with the situation. It is probably also dependent on the setting which way should be used.

4.4.4 Other Aspects

One fundamental issue when it comes to prioritization of the questions is whether it makes sense to do it. It can be argued that all questions contribute to their goal, and thus should not be considered separately (i.e., prioritization should only be done at goal level). However, it could also be argued that a set of questions for a goal is hardly ever complete, correct, or verifiable [15], given that it is not possible to do an exhaustive "search" for questions during the workshop. Furthermore, it is not necessarily so that all questions need to be answered in order to fulfill the goal; perhaps a subset of the questions, when answered, enable partial fulfillment of the goal (the most important part of the goal). For example, one goal may include questions that cover both process adherence and effectiveness. It may be so that the process adherence questions are of primary importance while effectiveness (in this particular goal) is of secondary importance. The prioritization of questions should be viewed as the task of finding a relevant subset of questions.

Another aspect of the prioritization that has been discussed earlier is weighting of stakeholders, i.e. different persons would influence the prioritization to different extents. One possible weighing scheme would be to let GQM-workshop participants have more influence, as they have more insight in the goals and questions. Another possibility is to give people influence based on their amount of involvement in the process at hand. Weighting of stakeholders has not been tried out during the case studies presented, and the implications of a particular weighing scheme with respect to the validity of the results must of course be considered carefully before using it.

5. CONCLUSIONS

In this paper, an extended version of the goal question metric (GQM) approach has been presented. GQM enables the definition of a measurement framework where metrics are connected to goals through questions. The benefit of this is that each metric serves a purpose, and no metrics are defined for the sake of measuring. However, three issues with the ordinary GQM approach that this extended approach deals with are:

1. The number of metrics can become quite large, which can be an obstacle to the industrial application of a measurement framework.
2. The resulting measurement framework may not cover all the relevant perspectives in relation to the process it will measure. Hence, the tree may be unbalanced or not complete.
3. Only the perspectives of those defining the GQM tree are taken into account, which may hamper the organizational acceptance.

The contribution of this paper is an approach that uses prioritization of goals and questions to deal with the first and third issues above, and categorization of questions to deal with the second. The prioritization is used to reduce the size of the measurement framework and will result in a partial GQM tree which is relevant for all involved stakeholders. Categorization on the question level

means that categories can be used to assess the completeness (or balance) of the GQM tree as well as making sure that not too much reduction of measurements are made.

The approach has been tested (and has matured) in two industrial case studies, and has been well received by the involved people and the organization. The way of defining the measurement framework was seen as very positive and beneficial in terms of producing relevant measurements (the result was considered as important to include by all affected persons spoken to), not at least since the participants got the opportunity to step back and reflect on different things. Furthermore, the prioritization was seen as both simple and powerful, and it provided a good basis for the consensus meetings. In these meetings, the participants were very active in discussing relevant strategies for limiting the number of metrics based on the prioritized goals and questions. They also showed great interest both for the result as such and for the future introduction of the measurement framework in their organization.

A number of experiences from testing the approach in the case studies have been discussed in detail. Some of the notable ones are the importance of moderating the GQM-workshop, the influence the workshop participants' mindset has on the efficiency of the work, the importance of controlling the focus in the prioritization, and paying attention to horizontal dependencies in the GQM tree (typically dealt with in the consensus meeting).

To sum up, the extended GQM approach addresses a number of important issues related to measurement frameworks in general and GQM in particular. Besides the positive aspects from a measurement point of view, the presented approach also has proven to be very helpful as a driver for process improvements. Based on experiences and feedback from the case studies, the approach seems promising and successful with respect to its purposes.

6. ACKNOWLEDGMENTS

This work was partly funded by the *Knowledge Foundation* in Sweden under a research grant for the project "Blekinge – Engineering Software Qualities" (http://www.bth.se/besq).

7. REFERENCES

[1] Berander, P., and Jönsson, P. Hierarchical Cumulative Voting (HCV) – Prioritization of Requirements in Hierarchies, to be published in *International Journal of Software Engineering and Knowledge Engineering (IJSEKE) – Special Issue on Requirements Engineering Decision Support, 16*, 5 (October 2006).

[2] Clark, B. Eight Secrets of Software Measurement, *IEEE Software, 19,* 5 (Sept/Oct 2002), 12-15.

[3] DePanfilis, S., Kitchenham, B., and Morfuni, N. Experiences Introducing a Measurement Program, *Information and Software Technology*, *39*, 11 (1997), 745-754,

[4] Fenton, N. E., and Pfleeger, S. L. *Software Metrics – A Rigorous and Practical Approach, 2nd Edition,* PWA Publishing Company, Boston, MA, 1997.

[5] Hetzel, B. *Making Software Measurement Work – Building an Effective Measurement Program*, John Wiley & Sons, New York, NY, 1993.

[6] ISO/IEC 15939:2002. Software Engineering – Software Measurement Process, 2002.

[7] Kotonya G., and Sommerville, I. *Requirements Engineering – Processes and Techniques*, John Wiley & Sons, Chichester, England, 1998.

[8] Mendonça, M. G., and Basili, V. R. Validation of an Approach for Improving Existing Measurement Frameworks, *IEEE Transactions on Software Engineering, 26*, 6, (June 2000), 484-500.

[9] Park, R. E., Goethert, W. B., and Florac, W. A. *Goal-Driven Software Measurement – A Guidebook,* Handbook CMU/SEI-96-HB-002, Software Engineering Institute, Carnegie Mellon University, Hanscom, MA, 1996.

[10] Shepperd, M. *Foundations of Software Measurement*, Prentice Hall, Hemel Hempstead, England, 1995.

[11] Van Aken, E. M, and Coleman, G. D. Building Better Measurement, *Industrial Management*, *44*, 4 (Jul/Aug 2002), 28-33.

[12] Van Solingen R., and Berghout, E. *The Goal/Question/Metric Method – A Practical Guide for Quality Improvement of Software Development*, McGraw-Hill Publishing Company, Maidenhead, England, 1999.

[13] Van Solingen, R., and Berghout, E. Integrating Goal-Oriented Measurement in Industrial Software Engineering: Industrial Experiences with and Additions to the Goal/Question/Metric Method (GQM), *Proceedings of the Seventh International Software Metrics Symposium (METRICS 2001)*, (London, England, April 4-6, 2001), IEEE Computer Society, Los Alamitos, CA, 2001, 246-258.

[14] Van Solingen, R., Basili, V., Caldiera, Gianluigi, and Rombach, D. H. Goal Question Metric (GQM) Approach, *Encyclopedia of Software Engineering* (Marciniak, J.J. ed.), online version @ Wiley Interscience, John Wiley & Sons, 2002.

[15] Wang, Y., and He, Q. A Practical Methodology for Measurement Deployment in GQM, *Proceedings of the Canadian Conference on Electrical and Computer Engineering (IEEE CCECE 2003)* (Montreal, Canada, May 4-7, 2003), IEEE Computer Society, Los Alamitos, CA, 1329-1332 (vol. 2), 2003.