METHOD FOR OBTAINING CORRECT METRICS

Coral Calero, Mario Piattini, Marcela Genero

E.S. Informática-University of Castilla-La Mancha
Ronda Calatrava, 5
13071 Ciudad Real (Spain)
e-mail: ccalero@inf-cr.uclm.es , mpiattin@inf-cr.uclm.es , mgenero@inf-cr.uclm.es

Abstract: Metrics can be used as a mechanism for assuring product quality. However, metrics will have this application only if they are well-defined. To obtain correct metrics a number of steps must be followed. In this paper we present the method we have designed for obtaining correct metrics. This method is composed of the metrics definition, formal validation and empirical validation of the metrics. After these steps we can know if a metric is or not correct. However, this information is not sufficient and we must be able to make some kind of interpretation regarding the value that a metric takes. For this reason, we have added the psychological explanation step to the method.

1. INTRODUCTION

Metrics are widely recognized as an effective means to understand, monitor, control, predict and improve software development and maintenance projects (Briand et al., 1996), and also for determining the best ways to help practitioners and researchers (Pfleeger, 1997). Software engineers have been putting forward huge quantities of metrics for software products, processes and resources (Melton, 1996; Fenton and Pfleeger, 1997). Unfortunately, the definition of metrics was made by using only the practitioners experience without carrying out any kind of tests with them.

However, this cannot longer be the case. If we want quality products we must use metrics, but we need well-defined metrics .Therefore, we think metrics definition must be carried out in a methodological way, following a number of steps which ensure the reliability of the proposed metrics. Figure 1 presents the method we apply for correct metrics definition.

In this figure we have four main activities: metrics definition, theoretical validation, empirical validation and psychological interpretation. As shown in figure 1, the process is evolutionary and iterative. As a result of the feedback, metrics could be redefined based on discarded theoretical or empirical validation or based on the psychological explanation.

Figure 1. Steps followed in the definition and validation of the database metrics

In the next sections we will discuss each of the steps of the method in depth. Section 2 presents how to make the proposal of metrics, section 3 explores some formal frameworks for making the theoretical validation. Section 4 presents the types of empirical validation we can make and how to make them. The principles in which the psychological explanation must to be made are presented in section 5. Conclusions and future work will come in the last section.

2. METRICS DEFINITION

The first step is the proposal of the metrics. Although it looks simple, it is an important one for ensuring that metrics are correctly defined. This definition is made taking into account the specific characteristics of the product we want to measure and the experience of product designers and users of these products. However if we want a methodological way for defining the metrics, we can use the GQM (Goal-Question-Metric) method proposed by Basili y Weiss (1984) and refined by Rombach (1990). The objective of this method is to define the metrics based on the goal we want to reach from the measurement. This method states that the measurement must be made with a concrete objective. GQM defines an objective, transforms this objective into a set of questions and defines those metrics which can give the information needed to answer these questions. In this way, the GQM method is based on the fact that each metric must be defined based on a top-down schema. The result of applying the GQM method is a three level model (figure 2): the conceptual level where the objectives are defined (Goal), the operational level where the questions are defined (Question) and the quantitative level where the metrics are defined (Metric).

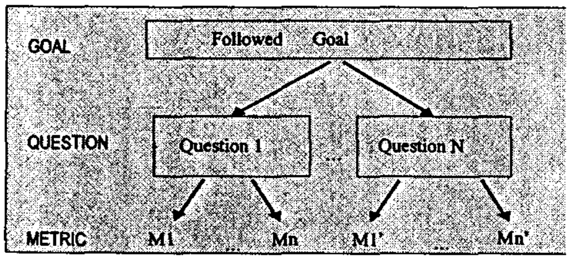

Figure 2. Three levels architecture of the GQM method

By using the GQM approach we can obtain metrics with a concrete goal but it does not ensure that the metrics obtained are correct. A complete application of the GQM method can be found in Van Solingen y Berghout (1999).

3. FORMAL VALIDATION

The second step is the formal validation of the metrics. Formal validation helps us to know when and how to apply the metrics. Unfortunately, as Van Den Berg and Van Den Broek (1996) point out there is nocommonly accepted standard for the theoretical validation of metrics but it is needed. However, there are two main tendencies in metrics validation: the frameworks based on axiomatic approaches and the ones based on the measurement theory. The goal of the property-based ones is merely definitional. On this kind of formal framework, a set of formal properties is defined for a given software attribute and it is possible to use this property set for classifying the proposed measures. The most well-known frameworks of this type are those proposed by Weyuker (1988), Briand et al. (1996) and Morasca and Briand (1997). The main goal of axiomatisation in software metrics research is the clarification of concepts to ensure that new metrics are in some sense valid. However, if we cannot ensure the validity of the set of axioms defined for a given software attribute, we cannot use it to validate metrics. It cannot be determined whether a measure that does not satisfy the axioms has failed because it is not a measure of the class defined by the set of axioms (e.g. complexity, length...) or because the axiom set is inappropriate. Since the goal of axiomatisation in software metrics research is primarily definitional, with the aim of providing a standard against which to validate software metrics, it is not so obvious that the risks outweigh the benefits (Kitchenham and Stell, 1997).

The measurement theory-based frameworks (such as Zuse 1998 or Withmire, 1998) specify a general framework in which measures should be defined. The strength of measurement theory is the formulation of empirical conditions from which we can derive hypothesis of reality. The measurement theory gives clear definitions of terminology, a sound basis of software measures, criteria for experimentation, conditions for validation of software measures, foundations of prediction models, empirical properties of software measures, and criteria for measurement scales which are necessary for knowing, for example, what statistical operations can be done with the metrics.

4. EMPIRICAL VALIDATION

The aim of this step is to prove the practical utility of the proposed metrics. Although there are various ways of performing this step, basically we can divide empirical validation into experimentation and case studies. Experimentation is usually carried out using controlled experiments and the case studies usually work with real data. Both of them are necessary, the controlled experiments for the initial approach and the case studies for enforcing the results . In both cases, the results are analyzed using either statistics tests or advanced techniques as C4.5 (a machine learning algorithm) and so on. Replication of the experiments is necessary because it is difficult to understand the applicability of isolated results from one study and, thus, to assess the true contribution to the field (Basili et al., 1999).The empirical study is necessary for testing and understanding the implications of the measures of our products. This can be done through hypothesis in the real world which must be proved with empirical data. The way to know if we must develop a controlled experiment or a case study depends on the level of control we have over the variables. If we have high control over the variables which can affect the hypothesis, we can develop an experiment but if, to the contrary, we cannot have a high control over the variables, it would be better to develop a case study. Another factor that should be taken into account when choosing between the two techniques is if the empirical study can be easily replicated . If it is easy to replicate it, it would be better to carry out a controlled experiment , if not, a case study is the best option. In table 1 these concepts are presented (Pfleeger, 1995). In the next sections we will present both techniques in more detail.

Factor	Controlled Experiments	Case Studies
Control level	High	Low
Difficulty of control	Low	High
Replication level	High	Low
Replication cost	Low	High

Table 1. Factors for selecting the best experimenttion technique

4.1. Experiments

The two characteristics which define an experiment are firstly that a completely controlled artificial situation is created and secondly that through an experiment we are aiming to detect a causal relation. Therefore, we can define an experiment as the creation of a controlled situation with which we aim to detect the causal relation among different events. The properties of an experiment are:

- Construct validity. The degree to which independent and dependent variables measure the concepts they try to measure.
- Internal validity. The degree of security with which we can establish the cause of the variations. An experiment will have internal validity based on the level at which our controls allowed us to reject alternative interpretations of the results. In order to achieve internal validity we must control the variables, we must identify and control the hidden variables, we must control the special sources of error and we must avoid the effect derived from the practice. In summary, the internal validity is the degree with which we can establish the variation causes.
- External validity. It is the degree of generalization of the results. All the experiment must have external validity, but not all of them will have the same generalization power. Some factors which can influence the external validity are: The similarity of the variables to real situations selection of the independent variable levels or the subjects attitude.

An experiment has a set of steps, of which we can distinguish four: problem determination, hypothesis creation, testing the hypothesis and results analysis. All these steps link an initial situation (where we want to know something) with a final situation (where we have obtained new information and knowledge).

4.1.1 Problem determination

An experiment cannot be applied to any problem. In general, we can only use an experiment when we can define the problem operatively and in terms of a causality. Also, we must be able to control the situation as much as possible. For these reasons, the number of situations where we can carry out a controlled experiment is very low. Another characteristic to be considered are the replicas, because as Basili et al. (1999) said replicas are necessary. To replicate an experiment is to do it again to see if we obtain the same results. Although it may not seem important, it is vital in science because unless an experiment has been replicated the results cannot be considered definitive.

4.1.2 Hypothesis creation: the design

From the step defined in the previous section, we cannot begin with the empirical test. Previously it is necessary to redefine the problem creating the work hypothesis. A work hypothesis is a concrete way to formulate an aspect of the problem, which can be tested empirically. This hypothesis usually includes the relation that we believe exists among the variables and this relation must be a causal one. From the hypothesis we must prepare the controlled experiment, specifying the concrete and controlled conditions in which we are going to test the hypothesis. This work is named experiment design . These designs can vary depending on the kind of work hypothesis.

4.1.3 Hypothesis Testing

To test the hypothesis, we must put the experiment in practice, which means that we must perform the experiment. The execution of the experiment must be in accordance with all the factors previewed in the design step. This adaptation is necessary because, if not, we will be performing different experiments and the results would have no relation with the hypothesis we are trying to test.

Consequently, it is convenient to carry out a pilot experiment with a small set of people in order to establish if the experiment is well-defined independent of the results obtained from it. When the experiment is finished, we have a collection of data which corresponds to the measures of the dependent variable made. Usually, these data are not directly interpretable because the experiments are performed taking into account that we have the statistics at our disposal, and therefore we are not interested in testing the variations in the dependent variables but in knowing if the variations in the dependent variable are due to the variations in the independent variables. As a result, the data obtained are submitted to certain statistics operations from which we obtain other data which constitute the results of the experiment. This statistical work is named data analysis.In Pfleeger (1995) a study can be found about which statistical technique is the most appropriate depending on the characteristics the situation.

4.1.4 Results analysis

The numeric results of an experiment have no significance in themselves. They are simply an indication of how the variable dependent has varied in the experimental situation. To interpret the meaning of these results they must be related to the antecedents with which the experiment was designed. The immediate antecedent of an experiment is the hypothesis from which we have started.

4.1.5 Experiment replicas.

As we have said previously it is necessary to replicate the experiments. A fundamental strategy for enabling this replication is to create laboratory packages which contain all the information of an experiment, such as the experimental design, the artifacts, the processes used...These laboratory packages simplify the experiment replica (Basili et al., 1999). There are various types of replicas:

1. Replicas which do not vary the hypothesis. They do not vary the dependent variables of the original experiment nor the independent ones. The strict ones duplicate the original experiment and are necessary for increasing the reliability of the conclusions about the validity of the experiment. They are used to demonstrate if the results of the original experiment are repeatable. The replicas that modify the way in which the experiment is made. They try to increase our confidence in the experimental results by studying the same hypothesis but changing some details of the experiment.
2. Replicas which vary the hypothesis. Although these replicas vary some variables they remain at the same level of specificity as the original experiment. They can be: (1) Replicas that vary the independent variables. These kind of replicas are used for investigating which aspects of the process are important by varying systematically some independent variables and examining the results. (2) Replicas that vary the variables that are intrinsic to the object study. These replicas vary the way in which the effectivity is measured in order to try to understand what dimensions of which tasks are more important and (3) Replicas that vary the context variables in the environment in which the solution is evaluated. This kind of replica is used for establishing which aspects of this environment are important because they affect the research process results and they allow us to understand the external validity.
3. Replicas that extend the theory. This kind of replica helps us to determine the limits of a process effectivity by making changes in the processes, products and/or models of the context in order to see if the basic principles remain.

4.1.7 Ethical aspects

In Empirical Software Engineering, ethical aspects are not yet considered but it is really important to consider them (Singer and Vinsen, 2000). Among others, subject and organization confidentiality should be considered, in order to conceal some results which are not directly related with the results but that can affect the subjects who developed the experiment or if some relation exists among the subjects and the experiment developer. All these factors would be taken into account because they can affect the subject when developing the experiment and the resistance of the enterprises to provide its data.

4.2 Case studies

There are occasions when the investigator only observes what happens in a natural situation. He does not introduce any variable to test if it affects the subjects' conduct and neither does he assign the subjects randomly to different groups, he only observes. There are three main problems which the exclusive utilization of a controlled experiment cannot resolve:

- Problems related to the nature itself of the variables we want to study. This problem can be divided in two: variables in which the investigator is interested but which cannot be manipulated and the variables that can be manipulated but this manipulation in a experimental context can provoke some suspicion regarding the subjects.

The second problem involves three cases in which the experiment can be used but alone is not enough. The first case corresponds to the problems which due to their nature cannot be understood with exclusively experimental methodology. The second case is when the experimenter considers the experimental study only as a programming step with the objective of improving his understanding of the object of the study and the third case is when the results obtained from the experiments do not follow the direction for seen.

Lastly, the third problem refers to the ethical aspects related the development of the experiments .

In Pfleeger (1995) a study can be found about which statistical technique is the most appropriate depending on the characteristics of the situation.

4.3 Advanced data analysis techniques

In both cases, controlled experiments and case studies, it is necessary to use not only statistical techniques, but also advanced techniques to analyze the results. As Morasca y Ruhe (1999) point out there is a great necessity to integrate techniques for discovering information and the measurement in software engineering. We give priority to the use of Machine Learning (ML) algorithms for several reasons. One of them is that real-life software engineering data are incomplete, inexact, and often imprecise; in this context, ML could provide good solutions. ML is also fairly easy to understand and use. However, perhaps the greatest advantage of an ML algorithm –as a modeling technique- over statistical analysis lies in the fact that the interpretation of production rules is more straightforward and intelligible to human beings than principal components and patterns with numbers that represent their meaning. This is very important for us because we want to obtain information about what kind of relationship can exist between our metrics and understandability.

5. PSYCHOLOGICAL EXPLANATION

Ideally we should be able to explain the influence of the values of the metrics from a psychological point of view. Some authors, as Siau (1999), propose the use of cognitive psychology as a reference discipline in the engineering of methods and the studying of information modeling. In this sense, cognitive psychology theories such as the Adaptive Control of Thought (ACT, Anderson, 1983) could justify the influence of certain metrics in database understandability. The knowledge of the limitation of human information processing capacity could also be helpful in establishing a threshold in the metrics for assuring database quality.

6. CONCLUSIONS AND FUTURE WORK

Metrics definition must be done taking a number of steps into account. These steps are metrics definition, formal validation, empirical validation and the psychological explanation. Metrics definition must be carried out taking the specific characteristics of the product we want to measure and the experience of the product designers into account. However, it can also be done in a methodological way by using the GQM approach.

Formal validation of the metrics gives us some important mathematical information about the metrics. This information can be used to classify the metrics (in the case of a property-based framework) or to know the scale to which a metric pertains and consequently, the statistical operations that can be applied to this metric (if we work with measurement theory based frameworks).

Empirical validation is used to know if a metric will be useful in practice or not. There are two main ways to carry it out, by controlled experiments or by case studies.

Finally, the psychological explanation explains the influence of the values of the metrics from a psychological point of view. As a result of the application of this method, we can obtain useful metrics which can be used in practice.

ACKNOWLEDGEMENT

This research is part of the MANTICA project, partially supported by the CICYT and the European Union (CICYT-1FD97-0168)

REFERENCES

Anderson, J.R. (1983). *The Architecture of Cognition*. Cambridge. MA: Harvard Universitiy Press.

Basili, V.R. and Weiss, D. (1984). A methodology for Collecting Valid Software Engineering Data. *IEEE Transactions on Software Engineering*. SE-10. No. 6. pp.728-738.

Basili, V.R., Shull, F. and Lanubille, F. (1999). Building Knowledge through families of experiments. *IEEE Transactions on Software Engineering*. July/August. No. 4. pp. 456-473

Briand, L.C., Morasca, S. and Basili, V. (1996). Property-based software engineering measurement. *IEEE Transactions on Software Engineering*. 22(1). pp.68-85.

Fenton, N. and Pfleeger, S. L. (1997). *Software Metrics: A Rigorous Approach* 2nd. edition. London. Chapman & Hall.

Kitchenham, B.A. and Stell, J.G. (1997) "The danger of using axioms in software metrics". *IEE Proc.-Soft. Eng.*. Vol. 144. No. 5-6. pp 279-285.

Melton, A. (ed.) (1996). *Software Measurement*. London. International Thomson Computer Press

Morasca, S. and Briand, L.C. (1997). Towards a Theoretical Framework for measuring software attributes. *Proceeding of the Fourth International. Software Metrics Symposium*. pp. 119-126.

Morasca, S. and Ruhe, G. (1999). Guest editor's introduction: knowledge discovery from empirical software engineering data. *International Journal of Software Engineering and Knowledge Engineering*. Vol. 9. No.5. pp. 495-498.

Pfleeger S.L. (1995). Experimental design and analysis in software engineering. *Annals of Software Engineering*. JC Baltzer AG. Science Publishers. pp. 219-253.

Pfleeger, S. L. (1997). "Assessing Software Measurement". *IEEE Software*. March/April. pp. 25-26.

Rombach, H.D. (1990) Design measurement: some lessonslerned. *IEEE Software*. 7(3). pp.17-25

Singer, J and Vinson, N. (2000). Ethics and Empirical Studies of Software Enginnering. *Empirical Software Engineering*. Vol. 5. No.2. Junio. pp. 11-16.

Van Den Berg and Van Den Broek (1996). Axiomatic Validation in the Software Metric Development Process (Chapter 10). *Software Measurement*. A. Melton (ed.). (Thomson Computer Press.

Van Solingen, R. and Berghout, E. *The Goal/Question/Metric Method: A practical guide for quality improvement of software development*. (McGraw-Hill, 1999).

Weyuker, E.J. (1988). Evaluating software complexity measures. *IEEE Transactions on Software Engineering*. 14(9). pp.1357-1365.

Whitmire, S.A. (1997). *Object Oriented Design Measurement*. Ed. Wiley.

Zuse, H. (1998). *A Framework of Software Measuremen*. Berlin. Walter de Gruyter.