

AGGREGATION PROCESS WITH MULTIPLE EVIDENCE LEVELS FOR EXPERIMENTAL STUDIES IN SOFTWARE ENGINEERING

Enrique Fernández

Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA. Argentina
Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires. Argentina
enfernan@itba.edu.ar

Abstract

Current meta-analysis-based procedures for aggregating experimental studies borrowed from other branches of science have proved not to be suitable for real-world software engineering. This paper presents an alternative aggregation process to the standard. It is based on an aggregation strategy with multiple evidence levels. Each evidence level is linked to a specific aggregation technique which is assigned depending on the quality and quantity of identified experimental studies.

1. Introduction

The number of experiments run within the field of software engineering (SE) has been increasing significantly for several years. These experiments cover a wide range of topics, such as testing techniques performance, requirements elicitation, programming language performance, etc. The experiments output interesting information in each case. If the information is to be of any use, however, the results should be aggregated to be able to get findings backed by as many studies as possible.

There have been some attempts at experiment synthesis for SE, e.g. [1], [2], [3], [4], [5], [6]. But none of these efforts have borne the expected fruit. Also the results of informal combinations [4], [2], [5], [6] were limited. Additionally, the attempts at combinations with statistical techniques [1], [3] turned out to be impracticable because of the differences in the design and execution of the experiments run by different authors.

Systematic review (SR) has recently been proposed as a method for systematizing the aggregation of empirical studies in SE [7], [8].

A SR is a procedure that applies scientific strategies to increase the reliability of the process of gathering, critically assessing and aggregating relevant empirical studies about a topic (SR) [9]. SRs have recently started to be used in SE [10], [11], [12]. However, while SRs provide a working framework useful for gathering and, to a lesser extent, critically assessing

experiments, it falls down on results aggregation. The reason for this failure lies in the fact that SR uses meta-analysis as an aggregation strategy. Meta-analysis is a collective name that refers to a set of statistical methods that are useful for combining experimental results, provided certain constraints are met, such as there being a minimum number of properly gathered and uniform experiments [13]. SE is far from satisfying these constraints. For example, meta-analysis cannot be applied in [12] because the experiments are heterogeneous, while in [11] the problem was the non-existence of replications.

Therefore, a special-purpose aggregation procedure needs to be developed that considers the particular features of a not very mature research environment like SE. This procedure should be able to be applied when there are not very many replications and when studies are heterogeneous, etc. This procedure will make it possible to improve the level of evidence gathered about the performance of techniques and tools that are in everyday use in SE and which are often adopted or acquired because of the reputation of their sponsors. This will make a SE development environment more engineering like.

Section 2 of this article describes the state of the art. Section 3 describes the problems identified in this research stage. Section 4 defines the importance of the problem. Section 5 establishes the materials and methods to be used to achieve the goals. Section 6 presents an introduction to the first version of the proposed solution. Section 7 describes the strategy for validating the results. Finally, section 8 describes some of the findings.

2. State of the art

The aggregation of experimental studies has a lengthy history in disciplines like education or psychology [14]. Lately though, its development has been driven by the health sciences [15].

There are two types of methods for aggregating the results of experimental studies: interpretative and non-interpretative methods [16]. Interpretative methods, like narrative summary [17] or grounded theory [18], are characterized by the findings being generated

according to the personal judgement of the people who analyse the results [19]. The results that these methods yield are not very reliable, because they are highly dependent on the reviewer, and their use has gradually declined in favour of the more reliable non-interpretative methods.

There are several alternatives within the non-interpretative methods, like case surveys [20], vote counting [13] or comparative analysis [21]. However, the most sophisticated method of all is meta-analysis [22]. According to [23] meta-analysis is the statistical analysis of a series of individual studies for the purpose of integrating the results into a measured summary.

What we are looking for when we do a meta-analysis is a numerical result that is a representative summary of the results of the individual studies and is, therefore, an improvement on the individual estimates. Meta-analysis is now implemented by means of the effect size technique [13]. This technique is conceptually simple: the global effect estimator is calculated as a weighted mean of the effect estimators of the individual studies.

For meta-analysis to yield representative results of the studies it covers, it is necessary to check that the individual studies can be summarized and combined. This property goes by the name of *homogeneity* and is determined through *statistical heterogeneity*. Statistical heterogeneity is useful for identifying whether or not the difference in study results is due to a random error. Another point to check when conducting a meta-analysis is to determine how dependent the result is on a given set of studies. This can be estimated through a *sensitivity analysis*. Sensitivity analysis is useful for finding out whether or not the final result took into account all the studies.

A number of authors [7], [8] defend the use of meta-analysis in SE. However, as mentioned above, it is almost impossible to apply a meta-analysis-based aggregation strategy in SE today. The key obstacles to its application are:

- Shortage of experiments, replications and homogeneity among experiments [7], [11]. As a consequence it implies a loss of important precision in the results using the standard techniques of aggregation, that are applied in general to a considerable amount of studies.
- Shortage of application of standards for reporting experiments. For example, [24] do not publish variances and [25] do not even report the means of the experimental results. Under these circumstances it is impossible to apply meta-analysis.

- Wide-ranging internal quality. For example, although they deal with the same research topic, there is a big discrepancy between [26] and [25] as regards study conception and make-up. This means that the studies cannot be considered replications and therefore cannot be used for a process of meta-analysis-based aggregation. In actual fact if they were, the heterogeneity analysis would invalidate the results.
- Non-standardization of the response variables. For example, the studies [27] and [28] use different response variables to analyse the same aspects. This means that these experiments cannot be considered replications and therefore they will not be able to be joined to the aggregation process.

Apart from meta-analysis, this group contains other less sophisticated alternatives whose application is subject to fewer constraints: case surveys [20], vote counting [13] and comparative analysis [21]. Unlike meta-analysis, which has been thoroughly researched, the applicability limits of these techniques, except for vote counting [13], have not been studied, and their application to SE has been negligible to date [12], [29]. Consequently, although promising, these techniques should be studied at length before being put into routine use in SE.

3. Problem description

There is at present no aggregation method specifically adapted to the needs of SE. On the one hand, the most reliable methods, like meta-analysis, have constraints that limit their applicability. On the other, there is a set of methods, like case surveys or comparative analysis, which are potentially applicable. They have, however, never been used in SE, and it is, therefore, not known whether or not they are any good for the purposes of aggregation. The goal of this research project is to develop an aggregation method that can get as many pieces of knowledge as possible by combining a maximum number of studies, irrespective of their quality. The method should be able to work within the SE-specific limitations, i.e. few studies that deal with the same response variable or treatment; non-standardization of response variables; and reporting of a small number of statistical variables.

4. Importance of the problem

Running experiments at different sites (in laboratories and industry) outputs partial results about the capabilities and applicability conditions of the technologies. For example, one set of replications might have concluded that testing technique A is faster than B, whereas another different set of replications

might have found that there are no differences between these techniques. However, as there is not always a perfect match, and sometimes quite a breach between the results of replications, different individual results need to be combined to get pieces of knowledge that practitioners can use in routine practice.

The limits and boundaries of a technology can be defined by accumulating the results of different replications. Aggregation methods can confirm the partial experimental results and determine what the real effects are (e.g. A is better than B), and even estimate the extent and certainty of the effects. However, the non-existence of aggregation methods tailored to real-world SE limits empirical SE's potential for providing empirically consolidated pieces of knowledge for informed decision making on real development projects in industry.

5. Materials and method

To gather the knowledge associated with our research goal, we follow the classical research tradition [30], [31], [32], identifying methods and materials needed for developing our research project. They are as follows.

Materials:

We will use three types of materials to develop the aggregation method:

- a) SE aggregation techniques: meta-analysis [7], [8], and less formal techniques [4], [5], [11] are now being used in SE; these techniques will be reviewed and probably used during this research.
- b) Aggregation techniques in other disciplines: apart from the aggregation techniques now being used in SE, we will look at other techniques that have not yet been used in this discipline, such as response ratio [33] or case surveys [20], to mention just a couple of examples.
- c) Experimental studies: there are at present a lot of experimental studies with the widest variety of features. They will be used to define the aggregation techniques (which should be able to correct the failings of the studies) and validate the feasibility of their use.

Method:

The tasks to be carried out to develop this aggregation process will be:

- a) Identify alternative aggregation techniques to *effect size* through literature review and expert consultation.
- b) Analyse the applicability conditions of the located aggregation techniques, i.e. under which

conditions the technique is applicable. For example, should the number of reported experimental subjects be the same in each study or can it vary from one study to another?

- c) Analyse the feasibility of the response estimated by the different identified aggregation techniques, i.e., what is the expected error level?
- d) Propose a method for applying aggregation techniques depending on their feasibility and the constraints on their application.
- e) Propose a strategy for interpreting the results output by the different techniques. This will be linked to the reliability of the response yielded by the aggregation technique.

6. Proposed Solution

To solve the study aggregation problem, we propose a multilevel aggregation strategy in which the aggregation techniques are used complementarily rather than alternatively or exclusively. The most appropriate techniques will be applied depending on the number and quality of the located studies, and the will be results analysed jointly. For example, if there were ten experiments, four that could be aggregated by meta-analysis and six that could not, an alternative aggregation technique would be used for the experiments that could not be aggregated by meta-analysis, and then the results would be listed alongside the results of meta-analysis. This way we would get more than one level of evidence. The first level is linked to meta-analysis and the others to the alternative techniques, which will be less reliable.

The proposed strategy is divided into five steps, as shown in Figure 1:

- **Classify studies:** The objective of classifying studies is to be able to group the different studies depending on their quality, the published response variables and the analysed treatment types.
- **Determine aggregation techniques:** The aim of this step is to identify the aggregation techniques that should be applied depending on the quantity and quality of located studies. If no aggregation technique can be used, an alternative step could be advisable: Apply generalization strategies or Generate findings.
- **Apply generalization strategies:** The goal of the generalization strategy is to remedy problems related to the small number of replications. To do this, the studies have to be searched for common characteristics that can be used to place treatments and/or response variables in a group with a higher level of abstraction (more general). Even though these groups with a higher level of abstraction are

not really replications, they can, because of their likeness, be considered conceptual replications and, therefore, as studies that can be aggregated.

- **Aggregate studies:** This step will apply the different aggregation techniques that can be used to combine the results of the experimental studies. The criteria and recommendations established in step 2 (Determine Aggregation Techniques) are used to do this.
- **Generate findings:** The objective of this step is to generate as reliable a report as possible including these pieces of knowledge. The results analysis in this report will start with the most reliable results (output by meta-analysis) and end with the least reliable results (output by alternative techniques). This way, if the results are compatible (all the levels of evidence state that one treatment is better than another), we will have got a more robust finding than we would have done by applying the techniques on their own. But if the results are not compatible, an attempt should be made to determine whether there are any as yet unidentified random variables or to state the need to generate new experiments linked to the topic. The following sections describe each of the above steps in detail.

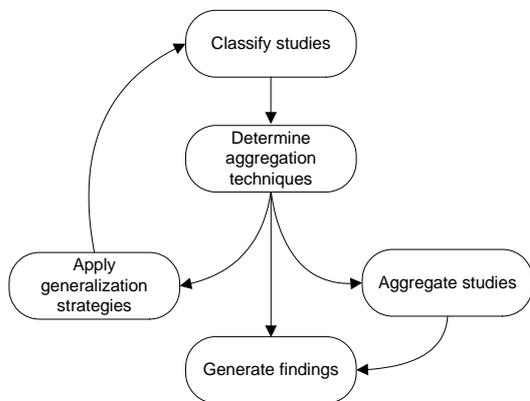


Figure 1. Aggregation process

6.1. Classify Studies

This research project is to examine a set of statistical techniques that, to assure the reliability of the response they estimate, call upon the empirical studies they analyse to meet a set of preconditions [12]. These preconditions are as follows:

- **Quality empirical studies:** a measure defining how well designed, executed and analysed the study has been. This provides an estimate of how reliable the results expressed in the study are. When the study is at risk of not being very reliable, a decision needs to be taken on whether

or not it should be part of the aggregation and, if so, to determine what aggregation techniques are going to be used. Although the quality of empirical studies is determined as a matter of course in most sciences, it is not completely defined within the field of SE. Therefore, this early version of the aggregation process will use the recommendations made in [8], and aims in future versions to apply the findings of on-going research by [34].

- **Complete reporting:** This is a second point to be analysed, as no matter how well built the study is, if the reporting does not cover a minimum set of parameters, the aggregation techniques will not be able to be applied. The key parameters are: means (M), variances (V) and number of experimental subjects (N). If the variances are not published, we need to find out whether the Student's t or Snedecor's F statistics have been published. Similarly, if the means are omitted, one remedy is to find out whether or not there were differences between them.
- **Representative treatments:** as there are not many study replications yet, this aggregation process proposes the application of a *generalization strategy* (see Apply Generalization Strategy step). This way, treatments that are not equal but are more alike than different can be placed in the same group. *Effect size* or *response ratio* estimation is not applicable because of the differences between these generalized studies. Briefly, the generalization of treatments places constraints on the aggregation types to be used. The same applies to response variable representativeness.

Based on the analysis of the above aspects, each empirical study should be assigned a category. This category determines what aggregation techniques can be applied to the study. Table 1 describes the main features of the studies in each category, as well as the technique that, for the moment, we consider applicable to each category.

Category	Study Characteristics	Aggregation Technique
1	This category admits experiments that are similar in terms of make-up and application domain.	Effect size
2	This category admits experiments that are similar in terms of make-up and application domain, but do not provide variances.	Effect size alternatives, like response ratio
3	This category admits experiments with slight reporting	Vote counting

Category	Study Characteristics	Aggregation Technique
	defects (only express differences between means or say that one treatment is better than another).	
4	Case studies and experiments <i>with generalizations</i>	Direct vote counting

Table 1. Description of study category

Table 2 can be used to deterministically assign studies a category. This table is an early version of an empirical studies classification and is likely to be updated. An obvious example is that the results on study quality that [34] is now investigating have yet to be aggregated with Table 2. Having categorized the studies, the classification needs to be further

Conditions				
Experimental Study Quality	Experiments	Experiments	Experiments	Case studies and experiments
Report publishes	Means, variances & subjects	Means, subjects	Means or express that one treatment is better than another	---
Treatments & Response Variables	None was generalized	None was generalized	None was generalized	<i>Generalization</i> was applied to experiment's treatments or response variables
Actions				
Assign Category	1	2	3	4

Table 2. Decision-making table for determining study category

6.2. Determine Aggregation Techniques

As mentioned earlier, this thesis proposes the use of a combination of different aggregation techniques. To do this, the number of studies linked to each treatment-variable pair has to be determined. This is because the applicability of aggregation techniques depends in part on the number of available studies [35].

If the quantity of category-1 studies is around 10, the classical meta-analysis based on weighted mean difference (WMD) [13] is applicable. If the precision of the mean effect is high, and no heterogeneity is present, the process can then be ended.

If the results are not satisfactory, or simply if we want to use as much studies as possible to try to obtain more solid results, we can use category-2 studies. The category-1 studies can be added to the category-2 studies and aggregation can be carried out using alternative aggregation techniques to WMD, as response ratios (RR) [36]. We think that such techniques can be used reliably with around 5 articles [35], but more research is needed.

In coherence with above, it is possible to get down in the category scale (to enlarge the number of available studies), but at the risk of obtaining less reliable results. To aggregate category-3 studies, we proceed the same way than before, and the vote-counting [13] looks like the most adequate technique to be applied. Finally, if there are category-4 studies, the studies of

decomposed depending on the response variables that the studies use. For example, suppose we want to determine which of two elicitation techniques, called A and B, is better. The empirical studies that these techniques test can use a range of response variables such as mean session time and quantity of acquired knowledge. As these variables are not compatible with each other, we have to decompose the set of available studies into: "Technique A versus Technique B using the Mean session time response variable" and "Technique A versus Technique B using the Quantity of information response variable". We will call this decomposition treatment-variable pair. This pair will guide the accounting process.

all four categories can added together, and the direct vote counting technique [11] could be applied. The above-mentioned techniques have been studied so far. However, we have not established the degree of reliability of the results of the aggregation in category-2, -3 and -4 studies. More research is needed in this topic, and the findings may still alter the proposed procedure.

6.3. Apply Generalization Strategies

The goal of generalization is to show up the common aspects of two treatments or response variables at a higher level of abstraction. The idea behind this process is to try to solve two problems that experiments run in SE tend to suffer from: small number of replications and non-standardization of response variables.

The better way to introduce generalization is by means of an example: Suppose that we wanted to find out whether C++ is better than its predecessor C, but there were only a couple of studies directly comparing these two programming languages. If apart from these two studies, there were some other studies comparing Delphi and Pascal, we can conjecture that, as C++ and Delphi are object-oriented languages and C and Pascal are structured programming languages, we will be able to get a more reliable finding by "lumping together" or generalizing C++ and Delphi (as well as C and Pascal),

because there are more studies available. Obviously, these findings do not answer the question of whether C++ is better than C, but they do answer a very similar question that yields light about the first question.

We think that the generalization strategy may be applied when there are not enough category-1, -2 and -3 studies to obtain reliable results using any aggregation technique. In this case, it seems reasonable to merge different (but not disparate; generalization should be applied with extreme care) treatments or response variables, and apply again the aggregation process, as shown before in Figure 1.

6.4. Aggregate Studies

In this step we will apply the different aggregation techniques that can be used to combine the results of the experimental studies. This will be done on the basis of the criteria and recommendations set out in the Determine Aggregation Technique step.

6.5. Generate Findings

Although this step is still under development, we are able to anticipate some details.

The different aggregation techniques used during the above steps of the aggregation process yield different results with different reliability levels. It is therefore necessary to analyse whether the results obtained by each technique are coherent with each other, which are more reliable and what discrepancies there are.

Additionally, the final report should contain two sections: "Generated Pieces of Knowledge" and "Possible Research Lines". The first section will describe the knowledge that has been gathered and for which there is firm supporting evidence. On the other hand, the second group will describe the conjectures made during data interpretation and the research questions that have not been able to be solved.

7. Validation Strategy

The reliability and versatility of this method will be validated by means of comparisons with other methods of aggregation. This will be done in two stages. The first stage will be run in a laboratory where a set of aggregation techniques will be applied to an artificially generated dataset and then the results will be compared. It will be analyzed the trustworthiness of the results of techniques, when the amount of experimental studies is low, and it will try to determine which variables may influence on the reliability of the results. The second stage will search for real systematic reviews in which meta-analysis has been applied and will compare the results of these reviews against the results output by the new aggregation process. This

validation will probably be run on studies conducted in other branches of science because of the constraints on applying a standard aggregation process to SE studies.

8. Conclusions and Future Work

The work done in this early stage of the research has revealed the complexity of aggregating empirical studies, as well as just how much work there is still to be done. Even so, the intended goal of developing an aggregation process especially tailored to SE has been proven to be feasible.

With respect to the following research steps, our priority is to continue analysing other branches of science to try to find more aggregation techniques and establish how precise they are. Shortly we intend to validate the proposed procedure using artificially generated data. This will be followed by a validation using real data.

9. References

- [1] Banker and Keremer; 1989; *Scale economies in new software development*. IEEE Transactions on Software Engineering. (15): 10, pp. 1199-1205.
- [2] Shull, F.; Carver, J.; Travassos, G. H.; Maldonado, J. C.; Conradi, R., and Basili, V. R.; 2003; *Replicated Studies: Building a Body of Knowledge about Software Reading Techniques*. Lecture Notes on Empirical Software Engineering. Chapter 2, pp. 39-84. World Scientific.
- [3] Hu, Q.; 1997; *Evaluating Alternative Software Production Function*. IEEE Transactions on Software Engineering. (23): 6, pp. 379-387.
- [4] Wohlin, C., Petersson, H., & Aurum, A.; 2003; *Combining data from reading experiments in software inspections: a feasibility study*. (pp. 85-132). World Scientific Publishing Co., Inc.
- [5] Juristo, N.; Moreno, A. M., and Vegas, S.; 2004; *Reviewing 25 Years of Testing Technique Experiments*. Journal of Empirical Software Engineering; 9(1 - 2):7-44.
- [6] Jørgensen, M.; 2004; *A Review of Studies on Expert Estimation of Software Development Effort*. Journal of Systems and Software. (70): 1-2, pp. 37-60.
- [7] Miller, J.; 2000; *Applying Meta-analytical Procedures to Software Engineering Experiments*. Journal of Systems and Software. (54): 1, pp. 29-39.
- [8] Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.
- [9] Goodman C.; 1996; *Literature Searching and Evidence Interpretation for Assessing Health Care Practices*; SBU; Stockholm.

- [10] Dyba, T., Kampenes, V., & Sjoberg, D.; 2006; A systematic review of statistical power in software engineering experiments. *Information and Software Technology*, 48(8), 745-755.
- [11] Davis, A.; Dieste O.; Hickey, A.; Juristo, N.; Moreno, A.; 2006; *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*; 14th IEEE International Requirements Engineering Conference (RE'06) pp. 179-188
- [12] Pickard, L. M.; Kitchenham, B. A., and Jones, P. W.; 1998; *Combining empirical results in software engineering*. *Information and Software Technology*; 40(14):811-821.
- [13] Hedges, L.; Olkin, I.; 1985; *Statistical methods for meta-analysis*. Academic Press.
- [14] Pillemer, D. and Light, R.; 1980; *Synthesizing outcomes: How to use research evidence from many studies*. Harvard Educational Review.
- [15] Evidence-Based Medicine Working Group; 1992; Evidence-based medicine. A new approach to teaching the practice of medicine. *JAMA*, 268(17), 2420-2425.
- [16] Noblit, G. W., & Hare, R. D.; 1988; *Meta-Ethnography: Synthesising Qualitative Studies*. Newbury Park, CA: Sage.
- [17] Fairbank L, O'Meara S, Renfrew MJ, Woodridge M, Sowden AJ, Lister-Sharp D.; 2000; *A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding*. *Health Technology Assessment*; 4: 1-171
- [18] Glaser BG, Strauss AL.; 1967; *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine de Gruyter.
- [19] Dixon-Woods, M.; Agarwal, S.; Jones, D.; Young, B., and Sutton, A.; 2005; *Synthesising qualitative and quantitative evidence: a review of possible methods*. *Journal of Health Services Research and Policy*. ; 10(1):45-53B(9).
- [20] Yin, R. K. and Heald, K. A.; 1975; *Using the Case Survey Method to Analyze Policy Studies*. *Administrative Science Quarterly*; 20(3):371-381.
- [21] Ragin, C.; 1987; *The comparative method: moving beyond qualitative and quantitative strategies*. Berkeley, California: University of California Press.
- [22] Straus, S. E.; Richardson, W. S.; Glasziou, P., and Haynes, R. B.; 2005; *Evidence-Based Medicine*. Churchill Livingstone.
- [23] Cochrane; 2007; *Curso Avanzado de Revisiones Sistemáticas*; www.cochrane.es/?q=es/node/198
- [24] Burton, A., Shadbolt, N., Rugg, G. y Hedgecock, A.; 1990. *The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise*. *Knowledge Acquisition* 2(2): 167-178.
- [25] Crandall Klein, B. y Asociados; 1989. *A Comparative Study of Think-Aloud and Critical Decision Knowledge Elicitation Method*. SIGAR Newsletter, April 1989, Number 108, Knowledge Acquisition Special Issue, pp. 144-146.
- [26] Burton, A., Shadbolt, N., Hedgecock, A. and Rugg, G.; 1988; *A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1*. Proceedings of Expert Systems '87 on Research and Development in Expert Systems IV, pp. 136-145.
- [27] Agarwal, R.; Tanniru, M.; 1990; *Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation*; *Journal of Management Information System*, M.E. Sharpe; Vol. 7 N. 1
- [28] Woody, J.; Will, R.; Blanton, J.; 1996; *Enhancing Knowledge Elicitation using the Cognitive Interview*; Expert system with application; Vol. 10 N. 1
- [29] Mohagheghi, P., & Conradi, R.; 2004; *Vote-Counting for Combining Quantitative Evidence from Empirical Studies - An Example*. Proceedings of the International Symposium on Empirical Software Engineering (ISESE'04).
- [30] Kumar, R.; 1996; *Research Methodology: A Step-by-Step Guide for Beginners*. Addison Wesley.
- [31] Creswell, J. 2003. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications.
- [32] Marczyk, G., DeMatteo, D., Festinger, D.; 2005; *Essentials of Research Design and Methodology (Essentials of Behavioral Science)*. John Wiley & Sons.
- [33] Gurevitch, J. and Hedges, L.V.; 2001; *Meta-analysis: Combining results of independent experiments*. *Design and Analysis of Ecological Experiments* (eds S.M. Scheiner and J. Gurevitch), pp. 347-369. Oxford University Press, Oxford.
- [34] Grimán Padua; 2007; *Propuesta de un proceso de revisión de estudios empíricos en Ingeniería del Software*; Internacional Doctoral Symposium on Empirical Software Engineering (IDoESE).
- [35] Lajeunesse, M; Forbes, M.; 2003; *Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques*. *Ecology Letters*, 6: 448-454.
- [36] Hedges; L.; Gurevitch, J.; Curtis, P.; 1999; *The Meta-Analysis of Response Ratios in Experimental Ecology*. *Ecology*. 80(4), pp. 1150-1156.