Updating a Systematic Review about Selection of Software Requirements Elicitation Techniques

Oscar Dieste*
Computer Science School.
Technical University of Madrid.
28660. Madrid. Spain
odieste@fi.upm.es

Marta Lopez
Computer Science School.
Complutensian University of Madrid.
28040. Madrid. Spain
mlf@fdi.ucm.es

Felicidad Ramos
INDRA Systems.
28850. Torrejon de Ardoz.
Madrid. Spain
framos@indra.es

Abstract

Quality of software products is closely related to the elicitation requirement process. Several studies point out that elicitation techniques achieve different results when applied in different contexts. This paper presents some recommendations about the situations in which elicitation techniques are useful. Recommendations are based on a previous systematic review, which was updated and expanded with 13 new empirical studies and more than 60 new empirical results. The aggregation process generated 5 new evidences and modified 4 existing ones. In the previous review, it was found that interviews were one of the most adequate techniques in most situations. The new evidence supports the same conclusion.

1. Introduction

Nowadays it is widely acknowledged within the software engineering community that requirements definition has a big impact on final product quality [10][40][38] Requirements Engineering (RE) is concerned with the elicitation, analysis, specification, validation and management of software requirements [39]. This paper focuses on the elicitation task and, more concretely, on the techniques applied to extract knowledge from the requirements stakeholders.

Although requirements elicitation appears to be a simple process in fact it is a really difficult one. Quite often, users do not know how to describe their tasks, may leave important information unstated, or may be unwilling or unable to cooperate [39]. Elicitation techniques aim to improve this communication process.

Despite the critical need for eliciting the right requirements, little research had been focused on identifying the most adequate elicitation techniques. Only ACRE [34] and recently the Unified Model of Requirements Elicitation [27][28] provide general frameworks. However, these works are by and large rooted on quite general theoretical foundations or expert opinion, leaving aside an increasingly large body of empirically-based knowledge.

Systematic Review (SR) is a technique employed in Evidence-Based Software Engineering (EBSE) [24], whose aim is to pool together the results obtained in different empirical studies and propose recommendations based on the best available evidence. In a previous work [16][21][17], the authors shown that SRs are a useful way to identify good practices regarding requirements elicitation. 30 different empirical studies were identified, reviewed and aggregated, generating 18 evidences about interviews, protocol analysis, sorting and laddering techniques.

A critical fact in any SR is the amount of evidence available. SR's conclusions are always based on the existing evidence when the SR is done, but as new empirical studies are discovered (because they were not identified before) or carried out, the conclusions of earlier SRs should be updated, either confirming or refuting the previous findings. This present paper updates the previous SR adding 13 new empirical studies and more than 60 empirical results. The subsequent aggregation process generated 5 new evidences and modified 4 existing ones. The new and modified evidences are in line with those in [17].

* Now the author is with Fraunhofer IESE, 67663 Kaiserslautern, Germany. Email: oscar.dieste@iese.fraunhofer.de

Generally speaking, they point out that interviews are the most effective elicitation technique in most situations, although its efficiency may be lower than some specialized techniques like laddering or card sorting in some cases.

This paper reports how the update of the previous SR was carried out and which evidences were obtained. It is structured as follows: Section 2 describes the motivation and introduces the respective background. Section 3 briefly describes the research methodology. The main findings are shown in section 4. Finally sections 5 and 6 present, respectively, the discussion and conclusions of this work.

2. Motivation and related work

There are a lot of elicitation techniques published in the literature either articles, such as [14][29] or books, like [32][33][41], among a lot of references. However, there is no agreement about how to select one (or several) techniques to work in a specific situation. In some cases, an insight into when an elicitation technique might or might not be applicable is provided [28]. But this is not enough because it is very limited. Maiden and Rugg have developed the most extensive research concerning the relationship between conditions and elicitation techniques [34]. ACRE (ACquisition of REquirements) is a faceted framework which includes categorization of techniques in relation to types of knowledge and types of communication. ACRE identifies six facets: purpose of requirements, knowledge types, internal filtering of knowledge, observable phenomena, acquisition context, and method interdependencies. Each of the facets can be interpreted as one or more questions to ask about an acquisition session. Answers to these questions inform method selection [34]. Therefore, the framework has been developed taking into account an analysis of the RE field and authors expertise to identify the critical factors for selection. However, the number of elicitation techniques analyzed was only 12. Since its development, ACRE had not been empirically validated; it is just a theoretical framework.

In a similar research line and with the same purpose of obtaining a model for technique elicitation selection, Hickey and Davis works were focused on assembling the collective wisdom of the most experienced analysts and experts. The final result was a ‘Unified Model of Requirements Elicitation’, which include a metaprocess of requirements elicitation and the selection of an appropriate technique [27][28].

The SR performed by Davis et al. pursues to find empirical evidence to make this selection [16]. Although it is important to consider the expert knowledge, the determination of the best practices in SE needs to be based on empirical proofs. There are a number of empirical studies on elicitation that have provided valuable results of which proper account has not been given in the literature. The aggregation of those empirical results can be used to confirm or refute some predefined and not proved assumptions about the effectiveness of requirements elicitation techniques. But also important is the maintenance of this aggregation over time, adding new references and gradually providing more pieces of knowledge to the RE field.

In the first SR researchers adapted Kitchenham’s procedures for performing systematic reviews and developed all the components and forms needed. For instance, the research question was specified as well as the search strategy and selection of primary studies, among others important tasks. Because there were no earlier systematic reviews in this domain, the search starting date was unlimited, whereas the ending date was set to March 2005 (inclusive). The empirical studies were identified by means of using a formal keyword search in well-known accepted databases, such as IEEEXPLORE, ACM DL and SCOPUS bibliographic databases as well as Google. Additionally, a review of the bibliographical references cited in the studies located by the above searches was also performed. Though extremely tedious, this later method was useful for identifying old studies not indexed in any database, studies with imprecise titles and grey literature. As a result, 564 candidate publications were identified.

The candidate publications were reviewed discarding all non-empirical studies. In the case of multiple publications, only one instance of each empirical study was considered. Admission and exclusion criteria (for instance, that studies were parallel trials) were used to further filter the candidate publications. The final result was a set of 26 publications reporting 30 empirical studies. The data from each study were extracted and entered in forms designed specifically for this particular study. A total of 43 elicitation techniques and 60 response variables were tested in the 30 selected empirical studies. After the aggregation process, a set of 35 aggregations were obtained. Those aggregations were the base for deriving 5 guidelines as a final result of this first SR.

3. Methodology

This work is an update of the SR described in [17]. That SR was carried out following the recommendations proposed in [31]. In this first SR, 53 publications containing potential empirical studies were identified. An initial search in online repositories and local library resources made possible to obtain 26 of those publications. These 26 publications contained 30 empirical studies which were reviewed and aggregated as already mentioned.

However, 27 potentially interesting publications were disregarded. As the conclusions of the SRs are contingent upon the available evidence, it was clear that a more thorough search (e.g. in international library services) was desirable. It made possible to obtain 13 out of those 27 publications. The other 14 publications (e.g. [11][25]) were considered impossible to locate, as they are quite old, grey literature. The overall literature flow is shown in Figure 1.

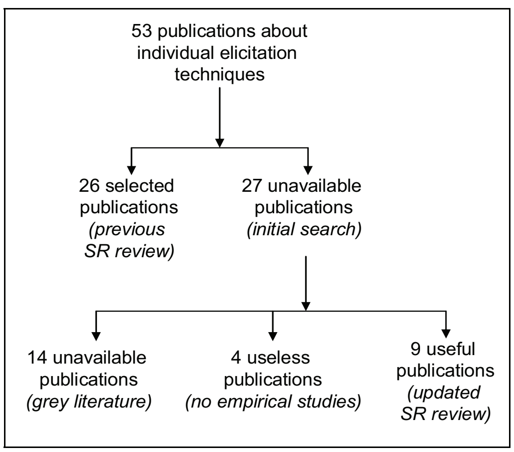

Figure 1. SR literature flowchart

Not all those 13 publications were useful. Four of them did not contain empirical studies at all or were papers published twice, so that they were discarded. The other nine ([1][2][3][4][5][6][7][8][9]) were useful and gave 13 empirical studies ([9] contained 2 different studies while [1] contained 3). The current updating work focused on those 13 empirical studies.

The tasks performed so far correspond to the initial stages of Kitchenham's procedure [23][30]. In a typical SR, both the review objective and the identification of studies would have been part of the SR itself. However, in this particular case, we are making an update of an existing SR (that is, the reference [17]) and therefore these tasks were obviously skipped.

The subsequent steps carried out during the updating work resemble closely Kitchenham's procedure [30], along with some modifications introduced in [17], as shown in Figure 2. This new process is described in [22].

However, the process was more difficult to perform than expected, because the update of a SR introduces problems unknown during the first execution. The most relevant problem was to relate the newly obtained empirical results and the previous ones.

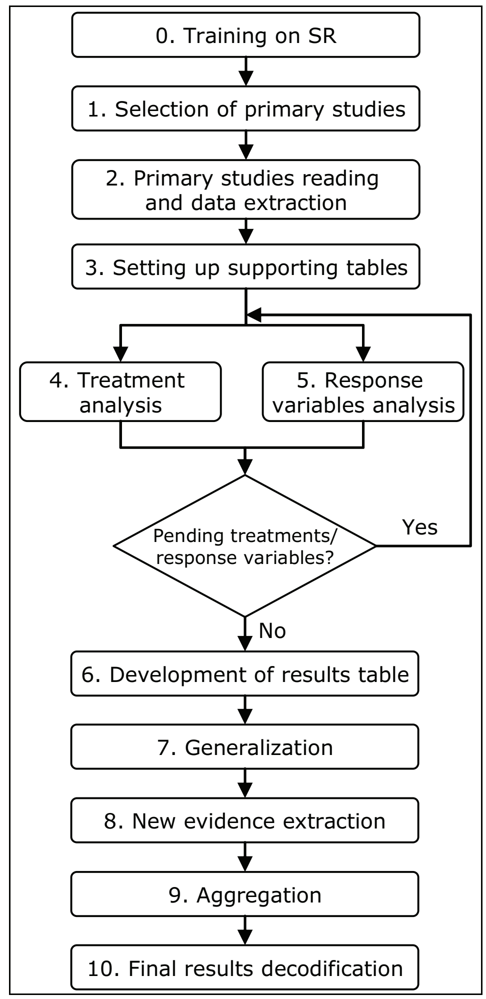

Figure 2. Systematic review process applied

Since we had no a glossary of terms from [17], the identification of the treatments (elicitation techniques) and response variables in the 13 new studies and the merge with the first SR's treatments and response variables was a complex task. We realized that the same technique could be named differently in diverse studies, although being the same, because the names are a subjective feature, depending on each author. Likewise, response variables suffered the same problem, aggravated by the fact that not only the name, but also the measurement procedure could vary.

If treatments and response variables of the present and past SR could not be merged, the updating work would be doomed to fail, because the combination of current and past empirical results would be unfeasible. Both SRs would be isolated efforts impossible to relate and, therefore, the number of potential evidences to obtain would be much lower. To solve this problem, it was necessary to catalogue the techniques and response variables tested in the empirical studies analyzed in [17], which implied to read the 26 initial publications besides the other 9 specific of this work. It represented a lot of effort which could be saved if such a glossary would have been constructed during the initial SR.

Apart of this drawback, it was possible to perform the SR with only minor difficulties. The chosen papers were reviewed using the same extraction form proposed in [19]. Once this form for each experiment was filled, the information extracted was coded to facilitate the later process of aggregation. Codification is necessary because the personal preferences of the researchers involved in the aggregation can bias the process. For instance, if the researchers believe in the superiority of one technique (i.e. interviews) over another (i.e. protocol analysis), then they may disregard the evidence contrary to their expectations (probably unconsciously). The codification assigns a code to each technique and response variable, making possible to aggregation results blind to the researchers involved. We used the same codification applied in the first SR in order to make results comparable and compatible [17].

After codification, the aggregation procedure was applied. In [31], the use of meta-analysis is recommended. However, when performing the previous review, we already realized that meta-analysis was of limited applicability. Meta-analysis requires: (1) two or more replications of the same experiment and (2) the reporting of some concrete statistical data (number of subjects per arm, means and variances). Such conditions do not usually hold in current SE experiments. In most cases, experimental replications are not exact, as they change the experimental task, type of subjects or experimental settings. On the other hand, SE experiments do not follow any reporting standards, and it often happens that variances or even means are not reported in the publications, making meta-analysis impossible.

Taking into account these characteristics, we developed in [17] an aggregation procedure based on an improvement of comparative analysis [37], quite similar to vote counting [26]. The key underlying idea is that while most empirical studies on elicitation techniques are not replications, many of them have similar factors and variables, and therefore some level of combination is possible. Concretely, if two studies test similar techniques and the outcomes are comparable, therefore a coincidence in the outcomes suggests the existence of a true effect.

An example is the best way to explain how the aggregation procedure works. Consider the studies [2] and [13]. They participate in the aggregation #12, but they are not replications of the same experiment (as they are, for example, the studies [2] and [8], or [12] and [15]) and therefore they cannot be aggregated by means of a meta-analysis. However, studies [2] and [13] have a lot of things in common.

On the one hand, the study [2] compares the duration of an elicitation session using interviews vs. a session using laddering, while [13] compares the duration of an elicitation session using free elicitation vs. a session using hierarchical dichotomization. Therefore, the response variable is the same.

Nevertheless, the techniques tested in both studies are different. Sure? A closer look to [13] reveals that "free elicitation" is the name given to interviews in marketing research (the field to which the paper [13] belongs), but it is not the case for laddering and hierarchical dichotomization. Definitively, they are different techniques and in consequence the studies [2] and [13] are not aggregable.

The only way to escape of this problem is to relax the conditions imposed to the studies to aggregate them together. Of course, we cannot mix apples and oranges, but some level of flexibility can be admitted. For example, laddering is essentially a technique with which the analyst composes n-ary hierarchies based on the information provided by the user. Hierarchical dichotomization is almost the same, with the only exception that the tree is binary instead of n-ary. Is that difference enough to make both techniques non-aggregable. Maybe, but maybe not. We can expect that techniques so similar behave in similar ways. Therefore, we can make the decision than laddering and hierarchical dichotomization are versions of the same technique, and that decision makes that [2] and [13] can be aggregated together. If both studies identify the same effect (for example, that interviews > {laddering | hierarchical dichotomization}), then our confidence in that effect increases, and vice versa.

4. Main findings

After performing the SR, we obtained more than 60 new empirical results. Those results, as well as all forms and supporting tables used, cannot be shown here but they will be published in [23]. Anyhow, that raw material does not have primary interest for the practitioner. The real interest lies in the combination of those empirical results among themselves, as well as with the results of previous SR. This combination or, more precisely, aggregation process, produces the evidences which can be later used to identify in which situations a given elicitation technique is useful. For details about how this aggregation process is performed, see [17] as well as [18].

Table 1 shows the results obtained after the aggregation process. The aggregations shown in the table are only those obtained during the updating work or those obtained in the previous SR but modified by the empirical results newly identified. A comprehensive table can be obtained from [23].

Table 1. Results of the aggregation

ID	KEY	Result of the aggregation	Support	Neutral	Opposes
12	REFUTES	There do not appear to be any differences in terms of session duration between unstructured interviews and laddering *(this evidence is not longer valid)*	[13]	[12][15]	[2]
16	REDUCES	Transcription time cannot be established as being longer for introspective techniques, like protocol analysis, than for unstructured interviews or vice versa	[3]	[12][15]	
21	REINFORCES	Transcription time cannot be established as being longer for sorting techniques than for laddering or vice versa		[3][13][12]	
33	REINFORCES	Laddering gathers fuller information than sorting techniques	[3][13]	[12]	
36	NEW	The efficiency of unstructured interviews is greater than scaling techniques	[2][8]		
37	NEW	Laddering and scaling techniques have the same efficiency	[2][8][13]		
38	NEW	Scaling techniques are more difficult to apply than unstructured interviews	[2][8]		
39	NEW	Laddering is more difficult to apply than unstructured interviews	[2][8]		
40	NEW	Laddering and scaling techniques have the same difficulty	[2][8]		

The first column of Table 1 contains the aggregation's ID (both for the previous and current SR). This ID is only used to ease the reference to aggregations. The second column contains one of the following codes:

- REFUTES: the newly gathered results refutes a previous aggregations.
- REDUCES:the new results cannot refute a previous aggregation, but reduces our confidence in it.
- REINFORCES: a previous aggregation is supported with new compatible results, and
- NEW: a new aggregation, not present in the existing set, has been identified.

Finally, the third, fourth and fifth columns are used to specify which studies provide evidence supporting the aggregation, which are neutral and which ones opposes to it (notice that references [1]-[9] were the ones analyzed in the present work; the others were analyzed in the first SR).

In some cases this type of table is not adequate to represent all types of evidences and its interpretation may be somehow difficult. For example, evidence 21 does not state a positive fact, such as "transcription time is longer for…", but they say "transcription time CANNOT BE ESTABLISHED AS being longer for…". In experimental terms, it means that no effect has been identified between techniques. In those cases "neutral" studies are really supporting the aggregation, and "support" and "opposes" studies deny it. This exception should be considered when reading aggregations 12, 16 and 21.

As shown in Table 1, the obtained aggregations are not supported by uniform empirical evidence. The same fact also happened in the first SR [17]. This fact caused the realization of a sensitivity analysis in order to know if the heterogeneity of the experimental studies, this is the contextual differences among experiments, could explain the conflicts among evidences.

The sensitivity analysis showed that the experiment performed in [12] was slightly different from others. Specifically, regarding the experience of the stakeholders of experiments, usually this experience could be define as high in all the experiments except for [12], where the experience must be define as low.

Therefore, due to this difference, the experiment [12] tended not to find differences among elicitation techniques (regarding response variables like "quantity of information") where other experiments did found differences. No other effects derived from contextual differences among experiments could be identified.

The sensitivity analysis had been repeated in our updating SR after joining the data obtained from both reviews (first and updating SRs). The final result is the same. In fact, we have identified that experiment [3] is quite similar to [12], with very analogous characteristics. However, it was not possible to corroborate the similarity of [3] and [12] due to the fact that [3] does not test response variables like "quantity of information".

We analyzed if the lacks of the experiment performed in [12] have an effect on the final aggregations. As shown in Table 1, [12] appears in the following aggregations: #12, #26, #31 and #33. The contribution of [12] in the four cases is a neutral evidence; this is, not identifying significant differences among techniques (as mentioned above). Fortunately, the impact of the experiment performed in [12] is very reduced because:

- The evidence provided by [12] is not needed to ground aggregations #12, #21, and #33. This is, if we do not consider the experiment performed in [12] in our aggregation process, the final findings would not change.
- The experiment performed in [12] is needed for aggregation #16. If we do not consider it, the aggregation would be refuted. Nevertheless, this fact would be positive, since our updating work already reduces the evidential level of aggregation #16. In other words, the interviews probably take less time to transcript than protocol analysis. In this sense we will consider aggregation #16 when elaborating recommendations.

Additionally, the generalizability of the obtained aggregations was also analyzed in [17]. The final result was that the heterogeneity of the primary studies, although high, did not seem to put the generality of the obtained aggregations in risk. On one hand, a certain degree of heterogeneity is desirable because the aggregations will be more valid when the supporting evidence is consistent across different experimental contexts [35][36]. On the other hand, the primary studies that support the aggregations are relatively uniforms within each aggregation, so much so that in many cases the aggregations are supported by studies which are quite nearest replications. This same conclusion maintains full validity in the context of the updating SR.

5. Discussion

In the previous systematic review we obtained several conclusions based on the available empirical evidence. Briefly, those conclusions were the following:

1. Interviews were the most effective elicitation technique. Overall, interviews performed better than sorting and protocol analysis, with very few exceptions. It was also observed that structured interviews performed better than unstructured ones, but it is not of interest for the present discussion.
2. Laddering is almost as good as interviews.
3. Sorting and protocol analysis can be considered low effective techniques. Overall, sorting is somehow better than protocol analysis, but the differences are of no consequence.
4. The effectiveness of interviews, laddering, sorting, and protocol analysis is comparable. We could not find differences among them either regarding elicitation or transcription time.

The newly obtained evidence points out in the same direction than the previously existing one. Concretely, the present study gathers more evidence for the effectiveness of interviews and laddering. On the one hand, aggregations #38 and #39 show that interviews are less difficult to apply than scaling and laddering. On the other, laddering is also found to provide more complete information than sorting (aggregation #33). Therefore, interviews appear to be the most effective technique overall, and laddering follows suit with the only drawback of being as difficult to apply as scaling techniques (aggregation #40).

The inferiority of sorting techniques is also corroborated by the present study, because the aggregation #33 (identified in the previous study) was reinforced with new evidence provided by [3]. In the case of protocol analysis, we could not find any new evidence for its inferiority regarding interviews and laddering, but neither against it.

The present study also confirms that the efficacy of interviews, laddering, sorting and protocol analysis is very similar to each other, or even the same. The aggregations #12, #16 and #21 show that there are not differences in the time needed to elicit or write down a transcription using any of those techniques.

As a final point, we did find new evidence about a new technique (scaling) not considered before. This technique (really, a set of different but highly related techniques) appears to be more difficult to apply than interviews (aggregation #38) and less efficient than interviews and laddering (aggregations #36 and #37). Therefore, scaling does not seem a serious competitor for interviews or laddering, but to have a clear idea of its potential, it would be interesting to know its behavior against sorting and protocol analysis as well.

To conclude, it is noticeable that only 3 of the 9 papers reviewed (that is, [2], [3] and [8]) have contributed to the set of aggregations shown in Table 1. The other papers did not contribute due to several reasons, such as: (1) their quality is quite low, making impossible to extract outcomes to aggregate and (2) they are too diverse, that is, the techniques and response variables that they test are very different to each other, making them inherently non-aggregable. If this fact were known beforehand, it would have been possible to reduce the effort of the review to than 1/3rd of the total (we would not have to read 6 out of 9 papers) achieving exactly the same results. However, this circumstance is not a defect of the present work. Quite the contrary, the same fact was already observed in the previous review, and it happens in almost every SR performed in SE. More details about this problem are given in [20], but the conclusion that we can extract here is fairly simple: to save time and effort when doing SR in SE, we have to focus on replications and closely related experiments. The other papers can be discarded with almost no risk of overlooking interesting results.

6. Conclusions

This paper presents the update of a SR work performed with the goal of identifying well-founded practices when selecting an elicitation requirements technique. There are a lot of arguments for supporting the fact that selecting the most accurate technique/s has a relevant influence upon the quality of the software product developed from the requirements gathered with it/them. Therefore, we would make software engineers aware of the need for guidelines derived from empirical evidence to select the most appropriate technique [21]. Until now they can use their own experience or expert's guidance, such as [28]. But we think that these recommendations should also consider the scattered knowledge of different empirical studies on individual elicitation techniques.

The most important result of our updating SR work shows that interview-based techniques seem to be the most effective elicitation techniques. We already established in [16][17] that interviews were able to extract more information that several other elicitation techniques, and also that the information extracted was more complete. In the present work, we could not find any contradictory evidence. Quite the contrary, the results corroborate the superiority of interviews regarding several dimensions, such as efficiency or difficulty of application.

We think that these findings need to be merged with those obtained from a SR focused on theoretical papers regarding individual elicitation techniques, for instance.

This will be one of our near future research lines. Besides this, in the long term we want to complete this elicitation techniques research with the study of group elicitation techniques, either from an empirical and theoretical perspective, and merge all the results obtained. Maybe then software community will be able to develop a comprehensive theory concerning the application of elicitation techniques, joining pieces of knowledge obtained from empirical, theoretical and expert knowledge perspectives.

7. References

References used in the updating work

[1] Bradburn, B., "A comparison of knowledge elicitation methods". International Conference on Engineering Design (ICED'91), pp. 298-305, 1991.

[2] Brandt, J. P. and Shook, S. R., "Attribute elicitation: Implications in the research context". *Wood and Fiber Science*, vol. 37, pp. 127-146, 2005.

[3] Burton, A. M., Shadbolt, N. R., Rugg, G., and Hedgecock, A. P., "Knowledge elicitation techniques in classification domains". Proceedings of the 8th Conference in Artificial Intelligence ECAI-88, 1988.

[4] Chao, C.-J. and Salvendy, G., "Impact of cognitive abilities of experts on the effectiveness of elicited knowledge". *Behaviour & Information Technology*, vol. 14, pp. 174-182, 1995.

[5] Grabowski, M., "Knowledge acquisition methodologies: Survey and empirical assessment". Proceedings of the Ninth International Conference on Information Systems, Minneapolis, MN, USA, pp. 47-54, 1988.

[6] Holsapple, C. W. and Raj, V. S., "An exploratory study of two KA methods". *Expert Systems*, vol. 11, pp. 77-87, 1994.

[7] Maiden, N. A. M. and Rugg, G., "Knowledge acquisition techniques for requirements engineering". Proceedings of the Workshop on Requirements Elicitation for System Specification, Keele, UK, 1994.

[8] Steenkamp, J.-B. E. M. and Van Trijp, H. C. M., "Attribute elicitation in marketing research: A comparison of three procedures". *Marketing Letters*, vol. 8, pp. 153-165, 1997.

[9] Wood, L. E., Davis, T. C., Clay, S. L., Ford, J. M., and Lammersen, S., "Evaluation of interviewing methods and mediating representations for knowledge acquisition". *International Journal of Expert Systems*, vol. 8, pp. 1-23, 1995.

Other references

[10] Boehm, B. W., McClean, R. K., and Urfrig, D. B., "Some experience with automated aids to the design of large-scale reliable software". *IEEE Transactions on Software Engineering*, vol. 1, pp. 125-133, Mar, 1975.

[11] Bont, C. J. P. M., "*Consumer Evaluation of Early Product-Concepts*". 1992. Delft University.

[12] Burton, A. M., Shadbolt, N. R., Hedgecock, A. P., and Rugg, G. "A formal evaluation of knowledge elicitation techniques for expert systems: Domain 1". In: "Research and development in expert systems IV". Proceedings of Expert Systems '87, the seventh annual Technical Conference of the British Computer Society Specialist Group on Expert Systems, ed. Moralee, D. S. Cambridge, UK: Cambridge University Press, 1987.

[13] Burton, A. M., Shadbolt, N. R., Rugg, G., and Hedgecock, A. P., "The efficacy of knowledge acquisition techniques: A comparison across domains and levels of expertise". *Knowledge Acquisition*, vol. 2, pp. 167-178, 1990.

[14] Byrd, T.A., K.L. Cossick, and R.W. Zmud, "A synthesis of research on requirements analysis and knowledge acquisition techniques". *MIS Quarterly*, Vol. 16, pp. 117-138, 1992.

[15] Corbridge, B., Rugg, G., Major, N. P., Shadbolt, N. R., and Burton, A. M., "Laddering - technique and tool use in knowledge acquisition". *Knowledge Acquisition*, vol. 6, pp. 315-341, 1994.

[16] Davis, A., Dieste, O., Hickey, A., Juristo, N., and Moreno, A. M., "Effectiveness of Requirements Elicitation Techniques: Empirical Results derived from a Systematic Review". Proceedings of the 14th IEEE International Conference on Requirements Engineering, Minneapolis, USA. September 11-15, 2006, pp.179-188.

[17] Dieste, O. and Juristo, N., "Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques". *IEEE Transactions on Software Engineering*, submitted, 2008.

[18] Dieste, O."TR4. Aggregation of empirical results". Available at http://grise.upm.es/2/, 2008.

[19] Dieste, O., "TR2. Systematic Review Protocol". Available at http://grise.upm.es/2/, 2008.

[20] Dieste, O., Grimán, A.C., and López, M. "Revisiones Sistemáticas: Recomendaciones para un Proceso Adecuado a la Ingeniería del Software". Proceedings of the 13th Conference on Software Engineering and Databases. Gijón, Spain. October 7-10, 2008.

[21] Dieste, O., Juristo, J., and Shull, F., "Understanding the Customer: What Do We Know about Requirements Elicitation Techniques?". *IEEE Software*, vol. 25, pp. 11-13, 2008.

[22] Dieste, O., López, M., and Ramos, F., "Formalizing a systematic review updating process". Proceedings of the 6th International Conference on Software Engineering Research, Management and Applications, Prague, Czech Republic. August 20-22, 2008.

[23] Dieste, O., López, M., and Ramos, F., "Formalizing a systematic review updating process". Universidad Complutense de Madrid, Technical Report, to appear.

[24] Dyba, T., Kitchenham, B. A., and Jorgensen, M., "Evidence-based software engineering for practitioners". *IEEE Software*, vol. 22, no. 1, pp. 58-65, 2005.

[25] Geiwitz, J., Kornell, J., and McCloskey, B. P., "An Expert System for the Selection of Knowledge Acquisition Techniques". Anacapa Sciences, Santa Barbara, CA, USA, Technical Report 785-2, 1990.

[26] Hedges, L.V. and Olkin, I. Statistical Methods for Meta-Analysis, Academic Press, 1985.

[27] Hickey, A. M. and Davis, A. M., "A unified model of requirements elicitation". *Journal of Management Information Systems*, vol. 20, pp. 65-85, 2004.

[28] Hickey, A. M. and Davis, A. M., "Elicitation technique selection: How do experts do it?". Proceedings of the Requirements Engineering Conference (RE'03), pp. 169-178, 2003.

[29] Hudlicka, E., "Requirements Elicitation with Indirect Knowledge Elicitation Techniques: Comparison of Three Methods". Proceedings of the 2nd International Conference on Requirements Engineering (ICRE'96), Colorado, USA, April 15-18, 1996, pp 4-11.

[30] Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., and Rosenberg, J., "Preliminary guidelines for empirical research in software engineering". *IEEE Transactions on Software Engineering*, vol. 28, pp. 721-734, 2002.

[31] Kitchenham, B.A. "Procedures for performing systematic reviews". Keele University TR/SE-0401, 2004.

[32] Lauesen, S., "*Software Requirements: Styles and Techniques*". Addison-Wesley, 2002.

[33] Leffingwell, D., and D. Widrig, "*Managing Software Requirements*". Addison-Wesley, 2000.

[34] Maiden, N. A. M. and Rugg, G., "ACRE: selecting methods for requirements acquisition". *Software Engineering Journal*, vol. 11, pp. 183-192, 1996.

[35] Noblit, G.W. and Hare, R.D., "*Meta-Ethnography: Synthesising Qualitative Studies*". Newbury Park, CA: Sage, 1988.

[36] Pfleeger, S. L., "Soup or Art? The role of evidential force in empirical software engineering". *IEEE Software,* vol. 22, no. 1, pp. 66-73, 2005.

[37] Ragin, C.C., "*The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*". Berkeley, CA: University of California Press, 1987.

[38] Standish Group, "The CHAOS Report". http://www.standishgroup.com/sample_research/PDFpages/chaos1994.pdf, vol. Oct 18, 2005.

[39] SWEBOK. Software Engineering Body of Knowledge. http://www.swebok.org, 2005.

[40] Tavolato, P. and Vincena, K. "A Prototyping Methodology and Its Tool". In: "Approaches to Prototyping". Ed. Budde, R. Berlin: Springer Verlag, 1984.

[41] Wiegers, K., "*Software Requirements*". Microsoft Press, 1999.