Identifying Moderator Variables Through Requirements Elicitation Experiments Limitations

Dante Carrizo
Universidad de Atacama
dante.carrizo@uda.cl

Oscar Dieste
Universidad Politécnica de Madrid
odieste@fi.upm.es

Marta López
Xunta de Galicia
marta.lopez.fernandez@xunta.es

ABSTRACT

Interviews are the most widely used elicitation technique in Requirements Engineering (RE). Despite its importance, research in interviews is quite limited, in particular from an experimental perspective. We have performed a series of experiments exploring the relative effectiveness of structured and unstructured interviews. This line of research has been active in Information Systems in the past years, so that our experiments can be aggregated together with existing ones to obtain guidelines for practice. Experimental aggregation is a demanding task. It requires not only a large number of experiments, but also considering the influence of the existing moderators. However, in the current state of the practice in RE, those moderators are unknown. We believe that analyzing the threats to validity in interviewing experiments may give insight about how to improve further replications and the corresponding aggregations. It is likely that this strategy may be applied in other Software Engineering areas as well.

Categories and Subject Descriptors

D.2.1 [Requirements/Specifications]

General Terms

Experimentation.

Keywords

Requirements Elicitation, Experimentation, Interviewing.

1. INTRODUCTION

RE, as a discipline within the development of systems and software, has been widely recognized as crucial [1]. Inadequate, incomplete, or ambiguous requirements have a critical impact on the quality of the software and the amount of rework to develop the final product [3]. Several authors highlight the need to focus on the requirements elicitation process to get the right requirements [1] [3] [10]. Interviewing techniques play a major role in RE elicitation since they are the techniques most widely applied [7]. Despite their relevance, little research has been performed about how to assess and improve interviews' effectiveness [7]. Empirical research in particular is really scarce, as there are only a few experiments in RE regarding interviews. This contrasts to other disciplines, like Psychology or Finances, where considerable research had been performed in order to analyze empirically interviews' efficiency, accurateness, influence of roles, etc.

Our work on experimentation applied to elicitation began with a systematic review on elicitation techniques [6]. We realized that there were some comparative experimental studies exploring the relative effectiveness of unstructured and structured interviews [2] [4] [11] [12]. Overall, structured interviews performed better than unstructured ones [6]. However, we believe that guidelines have to be defined as recommended by the Evidence-Based Software Engineering [9], and this implies the use of rigorous aggregation methods like meta-analysis.

Meta-analysis is a demanding method in terms of the number of required experiments, and therefore the existing experimental base ([2] [4] [11] [12]) is insufficient. We began to design and execute replications of those experiments to expand the existing datasets, adapting the particular types of interviews, response variables and experimental tasks to our context (laboratory experiment with students in Informatics). We performed a pilot study in 2006, and thereafter we executed experiments on a yearly basis (2009 excepted for logistic reasons): 2007 (analyzed and published [5]), 2008 (to be submitted), and 2010 (to be analyzed).

Designing replications is a challenging task, because the original experiments are not usually described at the necessary details to be repeated. Additionally, we need to explore the potential moderator variables to gain a deep understanding about the phenomenon of interest (interviews in our case), but those variables are largely unknown. To the best of our knowledge, the only working approach that addresses this problem right now is [8]. However, it requires that the replications are quite similar in order to analyze the possible influence of moderator variables. This is not applicable using the existing experimental base.

An alternative source of potential moderator variables can be obtained from the limitations and threats to validity analysis from each experiment.But since threats to validity are methodological restrictions, they can not be used to identify moderator variables. We are interested in the *limitations*. Though we realized that sometimes authors identify as limitations:

1) Aspects that threat the experiments' external validity (e.g.: convenience samples),
2) Characteristics of an applied technique (e.g.: variability on the application of a structured interviewing technique), sample, or process, and
3) Possible hypothesized reasons that explain the experimental results (e.g.: the influence of interviewees).

From an initial generic limit-threats set, only this last subset of aspects point to the existence of moderator variables, and that information may be used to design new replications. In this paper,

we present an exploratory example of the use of limitations to find potential moderator variables, using the interview experiments as examples. The description of the limitations needs for, at least, a brief overview of the experiments. However, due to space reasons, these outlines are not included here, although they are available at http://www.grise.upm.es/sites/extras/1/. The structure of this paper is as follows: section 2 presents the analysis of the experiment limit-threats; section 3 describes the identified moderators; finally, the conclusions are presented in section 4.

2. ANALYSIS OF EXPERIMENTS LIMIT-THREATS

Table 1 shows the limit-threats identified by Agarwal *et al.* [2], Browne *et al.* [4], Carrizo *et al.* [5], Marakas *et al.* [11], and Pitts *et al.* [12] (we use only the first author hereinafter, for short). In average, each experiment identifies 5-6 limit-threats, except Browne, that identifies only 2. The limit-threats have been classified according to the aspect they make reference. For example, Agarwa states that "*the experts constitute a convenience, rather than a random sample, even though random assignment was employed in assigning them to groups*". This limit-threat is referred to the sample used in the experimentation.

As shown in Table 1, all the limit-threats had been grouped according to the three main categories finally obtained: Process followed (how the experiment was carried out), Sample (the characteristics of the interviewers and interviewees), and Techniques (how the interviews were applied). These categories are quite general and probably they can be used to classify limit-threats in other areas (e.g.: testing).

Analyzing the limit-threats by category (rows), we can obtain coincidences, frequencies, etc. For instance, it can be noted that the main limit-threat identified for the Process category is that all of the experimentation processes are laboratory settings, with what everything this implies, as described Marakas. In all other cases, the experiments focus on particular concerns, like the number of sessions ([2]) or the complexity of the experimental tasks ([4]).

In all the five experiments the category most frequently cited as limit-threat is the Sample, which includes the interviewers, interviewees, coders or any other role needed to perform the experimentation. Some authors are concerned with specific sample problems, like its convenience character ([2]) or its motivation ([11]). However, all of the experiments agree in recognizing the experience and role-playing as limit-threats. Four of the works pointed out this issue focusing in the interviewer and only Carrizo in the interviewee.

The least relevant criteria, according to Table 1, is the technique issue. Focusing on the elicitation techniques applied in the interviews, Marakas and Carrizo present opposite perspectives, whether they accept or not the variability on the application of the structured techniques. This difference only shows two possible approaches to the experimentation, depending on the type of control of these elicitation techniques but it does not invalidate the experiment. They are just characteristics of those experimental designs. Other type of techniques are those used to represent the elicited data, as the DFD from Marakas, or those applied to code the data extracted from the elicitation technique, from Browne.

The next step was to analyze each limit-threat in Table 1 according to the following classification:

1) Real threats to the experiments' external validity; that is, aspects that prevented the experimental results to be extrapolated to more general populations. Since they are methodological restrictions, they cannot be used to identify moderator variables.
2) Characteristics, or any aspect which in fact is a feature of the process, sample, or technique. They can not be used for identifying moderator variables either.
3) Inklings, or possible hypothesized reasons for explaining experimental results.

According to the Merriam-Webster Dictionary, an inkling is defined as "*a vague idea or notion; slight understanding*". Hence, we use this term to denote a suggested notion which it is neither tested nor judged since it was not analyzed in the experiment. For instance, in Health we can find studies which demonstrate certain unintended logical connections within the scientific literature; this connections potentially reveal new knowledge or hidden hypotheses [13]. For instance, the *hidden* connection between the magnesium deficiency and the migraine headaches in medical journals which is only detected through text mining. This is an inkling: a neither experimented notion, nor an explicitly certitude, but a suggested and known notion which is mentioned in different experimental results. In this line, we are looking for inklings in our experimental base or, in fact, the sources of potential moderator variables.

Table 1. Number and description of *all limitations (or limit-threats)* per experiment

	Agarwal *et al.*		Browne *et al.*		Carrizo *et al.*		Marakas *et al.*		Pitts *et al.*		Total
Process	3	Only one problem domain No "gold standards" Only one session	-		1	Problems not complex	2	Lab setting Exploratory nature	1	Only one determination strategy	7
Sample	2	Convenience sample Role playing	1	Prior experiences of interviewers	2	Interviewee vs. problem No techniques predefined preferences of interviewees	3	Sample size Sample motivation Sample commitment	3	Interviewers' experience vs. problem domain Interviewers' quantified experience Only one interviewee	11
Techniques	-		1	Precision of the coding scheme	1	Variability of the technique applied	2	Minimum use of 70% of the time Only one representation method	2	Only one measurement of cognitive stopping rules Coding scheme focused on predefined taxonomy of requirements	6
Total	5		2		4		7		6		24

For instance, Marakas identifies as limit-threat that their study is an experiment in laboratory. Obviously, laboratory experiments are limited in the knowledge they can obtain for many reasons: strongly controlled environment, idealized settings, etc., but it does not imply the existence of any moderator. Likewise, Agarwal identifies as a limit-threat the fact that the subjects used in their experiment were a convenience sample. Similarly, this is a methodological restriction. An unrestricted sample would have to be used in an ideal situation. However, it does not point to the existence of any moderator variable either.

The limit-threat related with the number of sessions, number of problems, complexity of the problems, and numbers of techniques applied are related with the cost, effort and availability of the individuals involved. They are clear restrictions that affect the generalizability of the experimental results, but they do not imply the existence of any inkling that affects interview effectiveness (that is, a moderator). In the same line, the techniques 'limit-threat' from Carrizo, Marakas and Pitts can be addressed in the same way, since they describe particularities of those experiments that restrict the generalizability of results the same than above.

Table 2 presents the *inklings* left after purging Table 1 from the threats to validity. They can be denoted as inklings because they point to the lack of validity of the experimental results within each experiment's context (that is, internal validity). However, they are not design mistakes. For example, the experience of the interviewers may be one of the inklings because Browne points out that the experience may affect interviewer's effectiveness. Therefore, if experiments do not take measures to control the experience of the subjects, the results may be invalid. In other words: the interviewer's experience is a potential moderator.

Other examples are the '*precision of the coding scheme*' from Browne, or the '*coding scheme focused on predefined taxonomy of requirements*' from Pitts. They can be considered as inklings in the sense we are proposing because they can be a source of a measurement bias, which may affect to the analysis of the hypotheses. Also, the inklings grouped under the Sample criteria could be considered as potential source of biases, excepting the sample size issue, which is not considered an inkling but an influence factor on the statistical power. A larger sample size only increases the confidence of an estimate. All these selected inklings are considered in the following analysis.

3. IDENTIFICATION OF MODERATORS

In the experimental literature in other scientific fields (Finances, Health, etc.), the inklings that we have identified in Table 2 are usually related to specific types of biases. Bias, in this context, has the typical meaning of *Systematic Error*, pointing to undesired influences of diverse origin that need to be removed or minimized in order to increase experiments' accuracy.

It seems apparent that the authors of the experiments on interviews had a similar viewpoint about the inklings in Table 2 and that is why they listed them under the *Threats to Validity* sections of their works. In some cases, proceeding in such a way is completely justified, as the inkling is clearly a bias. For instance, the inklings listed under the *Techniques* category in Table 2 are instances of *measurement bias* (risk for the accurate determination of the values of the response variables). Other example is the *sample motivation*, which is an instance of *motivation bias*. Motivation is a prerequisite for performing adequately a task regardless of the field and it does not seem a legitimate research object. In both cases, they do not fit our purpose.

However, in many other cases this is not true, particularly in the Sample category, which is again the most populated one in Table 2. What may be a bias for some disciplines (as Finances), it may be a legitimate research object for RE. It is the case, to cite a clear example, with the experience of the subjects. It is not surprising that most of the inklings of this kind come out in the Sample category. In RE, particularly in interviewing, we are particularly concerned about the stakeholders, their particularities and the relations they establish with the problem under study. Therefore, those aspects are not biases or risks, but aspects that have to be considered in order to understand properly when and how interviews work. This is the reason why we realized that this kind of inklings were really showing us potential moderator variables.

The list below shows a classification of the inklings in Table 2, under the perspective of the potential bias that they may give origin to, according to the perspectives of other disciplines (Finances, Health, etc.). It is not an exhaustive list (we included only the most clear instances) but it is useful for a quick analysis:

- Artifact bias, related with the '*Interviewee vs. problem*' and '*Interviewer vs. problem domain*' inklings.

Table 2. Number and description of *inklings* per experiment

	Agarwal *et al.*		Browne *et al.*		Carrizo *et al.*		Marakas *et al.*		Pitts *et al.*		Total
Process	-	-	-		-	-	-	-	-	-	-
Sample	1	Role playing	1	Prior experiences of interviewers	2	Interviewee vs. problem No techniques predefined preferences of interviewees	2	Sample motivation Sample commitment	3	Interviewers' experience vs. problem domain Interviewers' quantified experience Only one interviewee	9
Techniques	-		1	Precision of the coding scheme	-	-	-	-	1	Coding scheme focused on predefined taxonomy of requirements	2
Total	1		2		2		2		4		11

- Interviewer bias, or any systematic error due to interviewer's subconscious or conscious gathering of data. Related with those inklings regarding *role playing* and *experience of interviewers*.
- Interviewee bias, related with *role playing* and, according to Table 1, with the possible identification *interviewee-problem* and the potential bias derived from the preference of use of one technique over another.

As explained before, the three items above cannot be considered biases from the perspective of interviews in RE because they are aspects that we need to know in order to explain the reasons for interview effectiveness and that we need to take into account to perform elicitation in practice. Therefore, from those inklings (or potential biases) we can identify the following moderators:

- Problem
- Experience
- Personal (psychological?) characteristics

And from those moderators, the following recommendations follow naturally:

- Perform interviews about different types of problems, of different size and complexity and, preferable, from different domains.
- Analysis of subjects' experiences. The more detail gathered, the best for controlling the experiment and obtaining higher quality data. For example, and wherever possible and appropriate, apart from the years, ask for the number and size of projects.
- Analysis of role assigned to each subject, based on his/her experience, aptitudes, knowing of the field, etc. Maybe introducing psychological tests or related measures to study subjects' personality could be useful. Do not forget that there are more roles than interviewers and interviewees and that they also can have influence on the final outcomes of the experimentation.

It is of course difficult to apply all these recommendations in practice due to the experiment specific characteristics, measures, lack of an appropriate pool of subjects, etc. However, these moderators may have an influence and have to be considered in interviewing experiment.

For different software engineering areas, moderators will surely differ (maybe not in the case of the subjects' experience). However, we think that a similar procedure applied to the interview experiments may work.

4. CONCLUSIONS

When reporting experiments, researchers tend to mix threats to validity and other limitations. However, it is interesting to differentiate them because methodological restrictions, like the type of sample or the numbers of subjects, do affect external validity. Other limitations or inklings, such as the expertise of interviewers (in the case of interview experiments), are not threats to validity.

Those inklings are, in reality, pieces of the theoretical background of the corresponding scientific area. For instance, the expertise of the interviewer may have an influence on the interview's effectiveness, as common sense suggests. However, dealing with theories is a complicated issue nowadays in Empirical Software Engineering. It is easier to think of them as possible moderator variables that may influence the experiments' results.

We believe that the analysis of the inklings identified in experiments may be a useful strategy to find moderators. The moderators thus found may be included in the design of new replications. In this paper, we have applied these ideas to an existing set of experiments about interviewing techniques. We are aware that our proposal does not have a rigorous and systematic formulation. We plan to improve it in the future.

5. REFERENCES

[1] Abran, A., Moore, J.W., Bourque, P., Dupuis, R., and Tripp, L.L. *Guide to the Software Engineering Body of Knowledge (SWEBOK)*. IEEE, 2004.

[2] Agarwal, R., and Tanniru, M. R. Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation. *Journal of Mng. Inf. Systems*, 7, 1 (Summer 1990), 123-140.

[3] Bell, T.E. and Thayer, T.A., Software Requirements: Are they really a problem?. *2nd International Conference on Software Engineering* (ICSE'76) (San Francisco, CA, 1976), 61-68.

[4] Browne, G. J., and Rogich, M. B.: An Empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques. *Journal of Manag. Inf. Systems*, 17, 4 (Spring 2001), 223-250.

[5] Carrizo, D., Dieste, O., Juristo, N., and Lopez, M. Estudio Experimental de la Efectividad de la Entrevista Abierta frente a la Entrevista Independiente de Contexto. *14th Workshop on Requirements Eng. (WER'11)* (Brazil, April 27-29, 2011), 41.

[6] Dieste, O., and Juristo, N. Systematic Review and Aggregation of Empirical Studies on Elicitation Techniques. *IEEE Transactions on Software Engineering*, 37, 2, (March/April 2011), 283-304.

[7] Hickey, A., Davis, A., and Kaiser, D. Requirements Elicitation Techniques: Analyzing the Gap Between Technology Availability and Technology Use. C*omparative Technology Transfer and Society 1*, 3 (Dec. 2003), 279-302.

[8] Juristo, N., and Vegas, S. Using Differences among Replications of Software Engineering Experiments to Gain Knowledge. 3rd Int. Symposium on Empirical Software Engineering and Measurement (ESEM'09) (Lake Buena Vista, Florida, USA, October 15-16, 2009). IEEE, 356-366.

[9] Kitchenham, B., Dybå, T., and Jørgensen, M. Evidence-based Software Engineering. *26th Int. Conference on Software Engineering (ICSE'04)* (Edinburgh, UK, May 23-28, 2004). IEEE Computer Society, Washington DC, USA, 2004, 273-281.

[10] Leuser, J., Porta, N., Bolz, A., and Raschke, A. Empirical Validation of a Requirements Engineering Process Guide. 13th Int .Conf. on Evaluation and Assessment in Software Engineering (EASE'09) (UK, April 20-21, 2009), 1-10.

[11] Marakas, G.M. and Elam, J.J. Semantic Structuring in Analyst Acquisition and Representation of Facts in Requirements Analysis. *Inform. Systems Research*, 9, 1 (March 1998), 37-63.

[12] Pitts, M.G., and Browne, G.J. Stopping Behavior of Systems Analysts During Information Requirements Elicitation. *Journal of Mng. Inf. Systems*, 21,1 (Summer 2004), 203-226.

[13] Swanson, D.R. Two Medical Literatures that are Logically but not Bibliographically Connected. *American Society for Information Science*, 38, 4 (July 1987), 228-233.