M. Höst, B. Regnell and C. Wohlin, "Using Students as Subjects - A Comparative Study of Students and Professionals in Lead-Time Impact Assessment", Empirical Software Engineering: An International Journal, Vol. 5, No. 3, pp. 201-214, 2000.

Using Students as Subjects – A Comparative Study of Students and Professionals in Lead-Time Impact Assessment

Martin Höst, Björn Regnell, Claes Wohlin
Department of Communication Systems
Lund University
P.O. Box. 118
SE-221 00 LUND, Sweden
(martin.host, bjorn.regnell, claes.wohlin)@telecom.lth.se

Abstract

In many studies in software engineering students are used instead of professional software developers, although the objective is to draw conclusions valid for professional software developers. This paper presents a study where the difference between the two groups is evaluated. People from the two groups have individually carried out a non-trivial software engineering judgement task involving the assessment of how ten different factors affect the lead-time of software development projects. It is found that the differences are only minor, and it is concluded that software engineering students may be used instead of professional software developers under certain conditions. These conditions are identified and described based on generally accepted criteria for validity evaluation of empirical studies.

1. Introduction

In this paper the performance of software engineering students in empirical studies is discussed and evaluated. There are two main reasons to that this issue is important. First, it is of course a measure of the quality of education. A large difference in performance between professional software developers in industry and students in their last year at the university may imply that the persons leaving the university need experience and training before they are able to fully contribute. A second reason for studying the differences between students and professionals, and the main motivation for this paper, is that many empirical studies in software engineering research rely on students as subjects, although attempting to generalize the findings to professional software developers. A representative example is the study on defect detection methods for software requirements inspection reported in [Porter95].

If there is a large difference between students and professionals, the validity of studies with students is threatened. On the other hand, if the students have a good knowledge of software engineering practices and their performance is comparable to that of professional developers, it may be argued that the validity of such studies is acceptable.

The reason to use students as subjects is often that they are available at universities and they are willing to participate in studies as part of courses they attend. In many cases it is possible to combine the learning objectives of the courses with the research objectives of the studies.

The performance of students can of course be measured and assessed in a number of different ways. In this study we measure it through their ability to point out which factors that affect the lead-time (or time to market) of projects. This is a task that is not trivial and it requires knowledge and understanding of professional software development. It is also a task that is important for software developers, but it is not taught in any courses, neither at the university nor in industry.

In Section 2, the usage of subjective factors is briefly explained and 10 subjective factors are presented. Section 3 compares the conception of students and professionals concerning the importance of these 10 factors with respect to the lead-time of projects. Section 4 summarises the conclusions of the study.

2. Subjective factors affecting lead time

One obvious way of comparing students and professionals is to let each group of subjects perform the same task and then measure the difference between the groups. The construct that is measured and compared in this study is based on the effect of subjective factors on the lead-time of projects. The task that the persons involved in the study were asked to perform is important in software engineering and non-trivial.

2.1. Subjective measures

It is not possible to capture all aspects of a project in objective terms. For example, in the early stages of a project, a risk assessment cannot be carried out solely based on measured data compared to experience data from earlier projects. In many cases there are no suitable data available, or the data is not reliable. This means that management has to rely on subjective data instead. This can, for example, be seen in [Moynihan97] where a number of project managers were interviewed (using the technique of personal construct elicitation) concerning factors that they consider as important situational constructs during project planning. 201 factors were found and grouped into 22 themes. Examples of themes are “Level of IT competence and experience of customer/users”, “Developer's familiarity with platform/environment/methods”, “Developer's previous experience with the application”, and “Maturity of the technology to be used”. It is not possible to measure objectively according to these themes in the early stages of a project. Instead we often estimate the risk using subjective measures based on experience. This type of estimation is an important task in the early stages of a project, although it is often not made explicit.

2.2. Ten factors affecting the lead time

This study is based on the factors presented in [Wohlin95]. Some of these factors are similar to the factors identified in [Moynihan97]. The 10 subjective factors are:

1. Competence, i.e. the competence of the development team.
2. Product complexity, i.e. the complexity of the product that is to be developed in the project.
3. Requirements stability, i.e. a measure of how stable the requirements are during the project.
4. Staff turnover, i.e. how frequently staff changes during the project.
5. Geographical distribution, i.e. how geographically distant the development teams are in the project.

6. Methods and tools, i.e. how advanced the methods and tools available to the developers are.
7. Time pressure, i.e. the time pressure perceived by the developers.
8. Information flow, i.e. how well information is spread in the project.
9. Priority, i.e., the priority given to the project by senior management.
10. Project management, i.e. how well the project is controlled.

The factors are further explained in [Wohlin95], where the effect of them on the lead-time of projects is evaluated. In 10 projects the project managers were asked to characterise the projects with respect to each factor. This information together with project reports and interviews were used to grade each subjective factor for each project on a scale from 1 to 5, where 1 corresponds to the worst alternative and 5 corresponds to the best alternative. For example, for geographical distribution, 1 corresponds to that the project is carried out by more than three organisations and 5 corresponds to that all work is carried out in one single organisation.

For the 10 projects the actual effort and lead-time was also measured. In order to obtain a normalised measure of the lead-time, the actual lead-time was divided by the required effort. That is, a project with a long lead-time compared to its required effort will obtain a high value, and a project with a short lead-time compared to its required effort will obtain a small value. In the rest of this paper the term lead-time denotes the normalised lead-time, which may be interpreted as the development speed.

The correlation between the lead-time and the grades obtained for each factor is displayed in Table 1. The values are taken from [Wohlin95].

Table 1. Correlation between factors and grades in 10 projects.

Factor	1	2	3	4	5	6	7	8	9	10
Corr	0.57	0.44	-0.71	-0.41	0.58	-0.58	-0.02	-0.65	-0.31	0.16

Since it is positive to obtain large grades but negative to obtain large lead-time, one would expect that the correlation values be negative. This is however only the case for factors 3, 4, 6, 7, 8, and 9. When correlation values are positive this may be harder to interpret. For example, the positive correlation between factor 1 and the lead-time indicates that the more competent people are in a project, the longer the lead-time will be. This is probably not the case. One potential explanation to the correlation may be that people with high competence are assigned to projects with a high risk of being late. In [Reel99] it is also claimed that "having too many stars creates ego issues and distractions, while not having enough can make the team struggling with small problems". It is, however, uncertain whether this is the underlying cause in this case. The actual cause of the positive correlation values is not the main focus of this paper, and therefore it is not further discussed here.

3. Studying the conception of the effect of the factors

In the study, the difference between people from industry and software engineering students has been investigated by letting people from the two groups carry out a non-trivial software engineering task individually. The participants have been asked to judge the effect of the 10 factors on the lead-time of projects by using a method denoted the Analytic Hierarchy Process (AHP) [Saaty80]. Based on this, it is possible

to compare the conception of students and professionals with the empirically observed effect of the factors (subsequently denoted *actual effect*) as described in Section 2.2, Table 1. The question concerning the difference between students and professionals may be divided into the following research questions:

- What is the conception of students and professionals?
- What is the difference between the conception of students and professionals?
- What is the difference between the actual effect and the conception of students and professionals?

In Section 3.1 the AHP method is presented and in Section 3.2 the operation of the study is presented with respect to how the two groups were involved in the study, time scheduling, etc. In Section 3.3, the results of the study are presented according to the research questions described above, and the validity of the results are discussed in Section 3.4.

3.1. The AHP method

The participating subjects were asked to express their conception of how the 10 factors affected lead-time by using the Analytical Hierarchy Process (AHP) [Saaty80]. The AHP method has been applied previously in a software engineering context, see e.g. [Karlsson97]. The method assumes n objects as input and requires that $n(n-1)/2$ comparisons are made, involving each possible pair of objects. Each comparison implies that two objects are compared on a ratio scale reflecting a certain criterion. The objects of comparison in this study are the 10 factors and the criterion is "effect on lead-time". AHP stipulates a scale from 1 to 9, and a typical comparison in this study may state e.g. that "factor A has very much larger effect on lead-time than factor B". The scale that was used is presented in Table 2.

Table 2. Interpretation of comparison scale.

Value	**Interpretation**
1	Of equal value - no difference
3	Small difference
5	Essential or strong difference
7	Very large difference
9	Extreme difference
2, 4, 6, 8	Intermediate values

AHP provides a matrix representation of the comparison results and an algorithm for calculating the resulting *rate* of each object on a ratio scale ranging from 0 to 1, where the sum of all rates equals 1. In this study the rates are interpreted as "the relative effect of each factor on lead-time". If, for example, factor A gets the rate of 0.14, this is interpreted as "factor A stands for 14% of the total effect on lead-time in relation to the other 9 factors". AHP also provides an algorithm for calculating a consistency ratio (CR), which indicates the level of consistency among the comparisons. In [Saaty80], it is recommended that CR should be 0.10 or less in order to be acceptable. A short description of how a comparison matrix is created and how

the calculations of the rates and CR are made can be found in [Karlsson97]. An in-depth explanation of the algorithms is given in [Saaty80].

The participants did not need to bother with the AHP calculation, as they were provided with a form where all comparisons where listed and the criterion and scale where explained. The participants needed only to assess which of two factors impacted lead-time most for each factor pair, and all AHP calculations where made by the authors during data analysis.

3.2. Operation

The students participated in the experiment in the spring of 1999. They were involved in three different sessions, where every person could choose which session to participate in. At every session the students first received some information about the study and the 10 subjective factors (about 10 minutes). Then they received one form each, where they should carry out pair-wise comparisons of the factors. Since there are 10 subjective factors, every person should carry out (10*9)/2=45 pair-wise comparisons. This took about half an hour. They did not receive exactly identical forms. They differed in the order in which the comparisons were carried out, and there were 8 different orderings. This approach was taken, because we wanted to reduce the effect of the order.

The professionals were involved in the study during the autumn of 1999. They were each given a document describing the study and the 10 subjective factors, and a form similar to the forms that the students were given. They then completed the forms independently and sent them to the authors when they were finished.

The difference between how students and professionals were involved in the study is due to practical reasons. It was possible to gather the students and present the study and letting them complete the forms. This was impossible for the professionals. However, we do not think that this should affect the study too much.

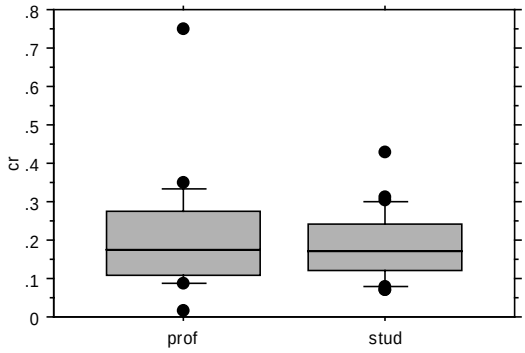

Figure 1. Box plots of the consistency ratios for each group of subjects.

In the study, 26 students and 18 professionals participated. Based on the consistency ratio one student (0.4283) and one professional (0.7512) was removed from the study. These persons showed too large inconsistencies. Since it concerns one person from each group and it only concerns 2 persons out of 26+18=44, the comparison is only marginally affected by this. That is, the presented results are based on 25 students and 17 professionals. The consistency ratios for all participants (including the subjects that were removed from the study) are shown in Figure 1.

3.3. Results

3.3.1. The conception of students and professionals

The results of the students are displayed in the left diagram in Figure 2, where the mean rate (i.e. the relative impact) of the factors is plotted (solid line). In the figure the standard deviation is plotted in two dashed curves representing *m+s* and *m-s*, where *m* is the mean value and *s* is the standard deviation.

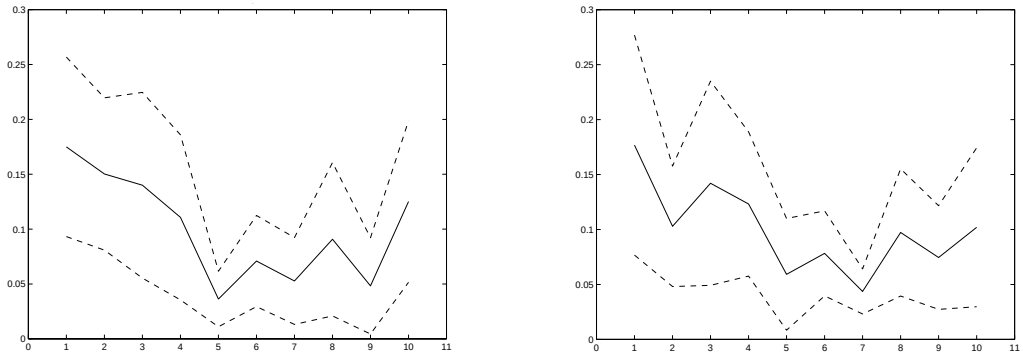

Figure 2. Results from AHP-analysis. The mean value is plotted for each subjective factor.

In the right diagram, the results from the professional developers are displayed in the same form as for the students. The mean values and their rankings are displayed in Table 3.

Table 3. Ranking of factors according to students and professionals.

Subjective factor	**students, rates**	**students, ranks**	**professionals, rates**	**professionals, ranks**
1	0.1750	1	0.1768	1
2	0.1502	2	0.1029	4
3	0.1400	3	0.1421	2
4	0.1107	5	0.1232	3
5	0.0363	10	0.0592	9
6	0.0709	7	0.0783	7
7	0.0527	8	0.0436	10
8	0.0908	6	0.0973	6
9	0.0482	9	0.0745	8
10	0.1252	4	0.1020	5

It can be seen that both students and professionals think that there is a difference between how important the factors are for the lead-time of a project.

The results are further elaborated below, where the difference between students and professionals are investigated.

3.3.2. The difference between the conception of students and professionals

If we look at Figure 2 and Table 3 we see that the overall conception of students and professionals are very similar. The Spearman correlation [Fenton96, Siegel88] between the mean values of students and professionals is 0.903 with a p-value of 0.0067, which corroborates the large correlation.

In order to investigate if there is any difference for a specific factor or the consistency, the difference between students and professionals has also been tested for each of these measures. This has been done with a Mann-Whitney U test with respect to consistency (CR) and rates (i.e. the relative impact) given to the 10 factors. The result is shown in Table 4.

Table 4. Result of Mann-Whitney tests for difference between subject groups.

Factor	CR	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
p	0.969	0.969	0.030	0.949	0.405	0.104	0.434	0.768	0.582	0.015	0.205

It can be seen that there is no significant (on the 0.10 level) difference for CR, F1, F3-F8, and F10. There is, however, a significant difference for F2 and F9 and there is almost a significant difference for F5. For these variables it can be concluded (from Figure 2):

- Factor 2: The students think that the product complexity is more important than the professionals do.
- Factor 5: The professionals think that the geographical distribution is more important than the students do.
- Factor 9: The students think that the priority given to the project by management is more important than the professionals do.

However, the differences are small and the diagrams in Figure 2 are very similar. It may be concluded that *there are only minor differences between the conception of students and professionals*.

3.3.3. The difference between the actual importance and the conception of students and professionals

In order to compare the conception of students and professionals to the actual importance of different factors, the conception has been compared to the correlation between the factors and the actual lead-time (i.e. the data presented in Table 1). This has been done in two different ways. First, we have looked at the Spearman correlation between the individuals' rates and the absolute values of the correlation values presented in Table 1. The absolute values were chosen since we have not asked people to judge the direction of the effect, just the magnitude. The Spearman correlation values of the individuals are plotted in the left two box plots in Figure 3.

Since some of the correlation values in Table 1 are hard to explain, we have also looked at the Spearman correlation between the priorities given by the subjects and the absolute values of the negative values in Table 1. The result is shown in the right two box plots in Figure 3. That is, the factors that are used in the correlation calculation are: 3, 4, 6, 7, 8, and 9.

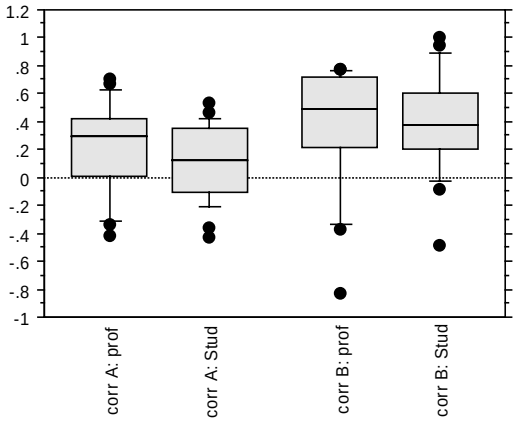

Figure 3. Box plots.

It can be seen that the correlation values are of the same magnitude for students and professionals. However, the professionals' correlation values are somewhat larger and the correlation values generally get larger if we only look at values from Table 1 that are negative, and hence easy to explain.

In order to investigate if there is any significant difference between students and professionals with respect to their correctness of their conception, the Spearman correlation values plotted in Figure 3 have been compare in a Mann-Whitney test. The test results in a p-value of 0.17 if all values in Table 1 are used (i.e. the left two box plots) and a p-value of 0.60 if values plotted in the right two boxes are used. That is, we can conclude that *there is no significant difference between the correctness of students and professionals*.

The study was performed with a non-trivial task, and the actual importance of each factor is, of course, not absolutely certain. It has been estimated based on correlation analysis as described above. The effect of the uncertainty of people and the uncertainty concerning the actual effect of the factors seems to be at least as large as the differences between the groups.

3.4. Validity

There are, of course, a number of threats to the validity of a study like this. Therefore it is important to analyse and evaluate the validity according to generally accepted criteria for this. In [Cook79, Wohlin99] four different types of validity are discussed: internal, external, construct, and conclusion. These are further discussed below.

3.4.1. Internal validity

Internal validity concerns the cause effect relationship, that is, if the measured effect is due to changes caused by the researcher or due to some other unknown cause. In this case it would mean that any measured difference between students and professionals is not due to that they are students and professionals. A number of potential threats to internal validity are listed and discussed below.

- **History:** If the environment has changed over time during the study this may cause a difference. The two groups were not involved at the same time. We have, however, not seen that this affects the result of the study.

- **Testing:** The study as such may result in a measured difference. For example, if one group in a study is faced to a pre-test, this may cause them to act differently in the study. The two groups received the same type of forms for the study. The only pre-test that was done in the study was a question to the professionals where they were asked to tell how many years they have been working in industry and in how many projects they have participated. We do not believe that this affects the outcome of the study.

  Another related difference between the groups is that the students received an oral presentation of the study and the 10 subjective factors, while the professionals received a written document. The students filled out the forms at a meeting and the professionals separately without a meeting. We do not believe that these issues are important for the outcome of the study.

- **Instrumentation:** If different groups are not measured in the same way this may cause a difference. We believe that we have handled this threat by letting people fill out similar forms.

- **Mortality:** If too many people do not complete the study, this may result in a difference. We do not believe that this is a major threat to the study.

- **Maturation:** Over time, the groups may be more interested in the study or they may be bored. Since people were only involved in the study for about half an hour this is not considered a major threat.

In the literature some additional threats are sometimes mentioned (see for example [Cook79, Wohlin99, Robson93]), but we do not believe that any of these are of any importance for this study. The threats discussed above contain the most often referred to threats, and we believe that we have control over the internal threats to the validity of the study. This means that when we have observed that the two groups are similar, we can say that they really are similar. There is probably no unknown third variable making two different groups look similar.

3.4.2. External validity

The external validity of a study describes the possibility to generalise its results. This is discussed in two dimensions below.

- **Selection of participants:** We can obviously not say that all possible students receive the same results as all possible professionals in all possible software engineering studies.

  The course from which the students were involved was a software engineering course for Masters students in the fourth year at the Computer Science and Engineering, and Electrical Engineering programmes at Lund University in 1999. We believe that it is possible to generalise this to students in the last years of a Masters programme in Computer Science or a related field. That is, they do probably not have to follow the course in 1999, study in Lund, or in Sweden at all. However, we cannot say that it is possible to generalise it to students in the first years of their education, or students in general.

  The mean value of the number of years that the professionals have been working in industry was about 11 years. This means that they are highly experienced and we are able to claim that the result is valid irrespective of the experience of the professionals. They work with real time embedded software, and we are not able

to say that the result would be valid for other types of applications. There is, however, nothing that says that it is not.

- **Nature of problem:** The problem is a non-trivial problem that the participants worked with for about half an hour. However, it involves a lot of knowledge about software engineering and it is not trivial, and as it was argued in Section 2.1, it is an important area in software engineering.

  It is uncertain to which other types of problems it is possible to generalise the findings. We think that the result at least is representative for problems that focus on project impact assessment. That is, the result is general for problems that concern situations when someone must estimate the impact of certain (subjective) factors on some project performance factor, such as the lead-time of the project. We also believe that the result is valid when it comes to peoples' general maturity and set of values in software engineering, and the understanding of dependencies and relationships in this area.

3.4.3. Construct validity

The construct validity reflects our ability to measure what we are interested in measuring. In this case the objective is to measure the difference between students and professionals when they participate in software engineering studies. Of course, this is related to the external validity with respect to the nature of the problem (see above). We cannot claim that the result is valid for all software engineering studies and experiments. The result is most likely valid for studies concerning the general understanding of dependencies and relationships in software engineering, for example project impact assessment tasks, as described in Section 3.4.2.

3.4.4. Conclusion validity

The conclusion validity describes our ability to draw statistically correct conclusions based on the measurements. For example, if we have very few data points and use a statistical test to compare two groups it is not possible to say that the two groups are equal. This is because the reason that we cannot find any difference may be that the test we apply has too low power when it is applied to the available data set, and not that the two groups are equal.

The data set that has been used in this study is small and it is possible that statistically significant results would be obtained if the data set was larger. However, the data set is not extremely small, and if we do not look at the significance levels but instead on the actual differences it can been seen that they are small. That is, even if we would use a much larger sample of subjects, if they would behave like the sample we have used, the differences would still be very small and probably of no practical importance. It can also be noted that the correlation between students and professionals described in Section 3.3.2 is large and significant. That is, it is not only that we cannot find any major difference between the two groups, we have identified a statistically significant similarity.

4. Conclusions

In the presented study students and professional software developers are compared in order to evaluate the validity of software engineering studies involving students instead of professional software developers. Only minor differences between students and professionals can be shown concerning their ability to perform relatively small tasks of judgement. The tasks that have been analysed are in the area of project impact assessment, which is related to risk assessment in the early phases of the life cycle. Regarding the conception of how lead-time is affected by the ten selected factors, the results show that:

- There are only minor differences between the conception of students and professionals.
- There is no significant difference between the correctness of students and professionals.

The external validity is based on the relevance of the performed tasks and the adequacy in the selection of subjects, and we argue that the result may be valid in the context of maturity and understanding of dependencies and relationships in software engineering. The result is probably valid for last year students at Masters programmes in Electrical Engineering and Computer Science and Engineering, and similar programmes, and for professionals working with software development for embedded systems.

The results do not contradict the assumption that last-year software engineering students are relevant as subjects in empirical software engineering research. However, this is only one limited study and it is important to replicate it in order to see if the same results are obtained or if other studies show contradictory results. It is also important to explore the generality of the results in further studies.

Another interesting opportunity in further research is to combine the results from several studies, which in combination include both students and professionals, e.g. using meta-analysis [Pickard98]. This opportunity may be exemplified by the study in [Porter95], which was carried out with students, and later replicated with similar results using professionals as subjects [Porter98]. This type of analysis will improve our understanding of empirical research with students as subjects, which in turn is a prerequisite for drawing general conclusions from research in academic environments.

Acknowledgements

The authors would like to express their gratitude to the people that participated in the study, both at the university and in industry. Part of this work is sponsored by The Swedish National Board for Industrial Development (NUTEK), grant 1K1P-97-09673.

References

[Cook79] T.D. Cook, D.T. Campbell, “Quasi-Experimentation - Design and Analysis Issues for Field Settings”, Houghton Mifflin Company, 1979.

[Fenton96] N. Fenton, S.L. Pfleeger, “Software Metrics - A Rigorous & Practical Approach”, Second Edition, International Thomson Computer Press, London, UK, 1996.

[Karlsson97] J. Karlsson, K. Ryan, “A Cost-Value Approach for Prioritizing Requirements”, IEEE Software, September/October 1997.

[Moynihan97] T. Moynihan, “How Experienced Project Managers Assess Risk”, IEEE Software, pp. 35-41, May/June 1997.

[Pickard98] L.M. Pickard, B.A. Kitchenham, P.W. Jones, “Combining Empirical Results in Software Engineering”, Information and Software Technology, 40(11), pp. 811-821, 1998.

[Porter95] A. Porter, L. Votta, V.R. Basili, “Comparing Detection Methods for Software Requirements Inspection: A Replicated Experiment”, IEEE Transactions on Software Engineering, 21(6), pp. 563-575, 1995.

[Porter98] A. Porter, L. Votta, “Comparing Detection Methods for Software Requirements Inspection: A Replication Using Professional Subjects”, Empirical Software Engineering, 3(4), pp. 355-380, 1998.

[Reel99] J.S. Reel, “Critical Success Factors in Software Projects”, IEEE Software, pp. 18-23, May/June 1999.

[Robson93] C. Robson, “Real World Research”, Blackwell Publishers Ltd., UK, 1993.

[Saaty80] T. Saaty, “The Analytic Hierarchy Process”, McGraw-Hill, 1980.

[Siegel88] S. Siegel, N.J. Castellan, “Nonparametric Statistics for the Behavioral Sciences”, McGraw-Hill, 1988.

[Wohlin95] C. Wohlin, M. Ahlgren, “Soft Factors and their Impact on Time to Market”, Software Quality Journal, No. 4, pp. 189-205, 1995.

[Wohlin99] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, A. Wesslén, “Experimentation in Software Engineering: An Introduction”, Kluwer Academic Publishers, 1999.